Preface

This volume contains the proceedings of the 4th workshop on 'Dynamic Perception' which was held on November 14 - 15, 2002 at the Ruhr-University of Bochum. The workshop focussed in an interdisciplinary manner on dynamic aspects of biological and machine perception, presenting and discussing recent work in this area. Special emphasis was on the promotion of scientific exchange between computer science (neurocomputing and artificial intelligence), psychology, and the neurosciences.

Specific topics on

- multimodal integration
- human movement analysis
- action and perception
- dynamic visual scenes
- optic flow
- gestalt laws and statistics
- cognitive influences on visual processing
- recognition and matching

were presented in 18 contributed talks and 32 posters. Invited talks were given by Jan-Olof Eklundh (Stockholm, SE) and William Phillips (Stirling, UK).

Looking at the development of the workshop series one must note that its scope has become more international. Of the 53 articles in this volume two thirds have first authors from Germany, but contributions also came from the UK (4), US (3), Japan (2), Spain (2), Sweden (2), and one each from Australia, Israel, Italy, Greece, and the Netherlands. We consider this a clear indication of the success of the workshop series.

The workshop was organized by section 1.0.4 (Image Understanding) of the German Society for Computer Science (GI) and supported by the EC research networks MUHCI (Multimodal Human-Computer Interfaces) and ECOVISION (Early Cognitive Vision). We specially thank the MUHCI consortium for additonal financial support and the Ruhr-University for providing the conference facilities and a generous contribution to the publication costs.

Our thanks also go to the contributors, whose high-quality abstracts made the inevitable selection rather difficult, and for their readiness to submit their final contributions in camera-ready form adhering to the layout requirements. We also thank the members of the program committee listed on the following page, Peer Schmidt for software help with formatting the final volume layout, Achim Schäfer for double-checking it, and Uta Schwalm for organizing the finances and the many details required for smooth operation of the workshop.

September 2002

Rolf P. Würtz
Markus Lappe

Organization

Conference Chairs

Rolf P. Würtz	University of Bochum
Markus Lappe	University of Münster

Financial Chair

Uta Schwalm	University of Bochum

Program Committee

Uwe Ilg	University of Tübingen
Christoph v.d.Malsburg	University of Bochum
Bärbel Mertsching	University of Hamburg
Jochen Müsseler	MPI for Psychological Research, Munich
Heiko Neumann	University of Ulm
Ioannis Pitas	University of Thessaloniki, Greece
Gerhard Sagerer	University of Bielefeld

Contents

Dynamic visual scenes

Cognitive influences on visual processing

Human movement analysis

Action and perception

Gestalt laws and statistics

Recognition and matching

Multimodal integration

Invited talks

The search for coherence through dynamic grouping and contextual modulation

William A. Phillips

Centre for Cognitive and Computational Neuroscience
Depts. or Psychology and Computing Science
Stirling University
Stirling FK9 4LA
Scotland, UK
email: w.a.phillips@stir.ac.uk

Cognitive neuroscience is dominated by evidence for semantic specialization. Different regions and different cells within regions process information about different things. We now need to understand how these diverse activities are coordinated. Coordination is necessary to enhance activity relevant to the current context, to combat noise, to make coherent choices, and to group activity into coherent subsets. The concept of Coherent Infomax formalizes this view within a theory of cortical computation. I will summarize evidence that coordinating interactions are implemented by a distinct family of physiological mechanisms that include synapses formed by NMDA receptors. Psychophysical studies of the effects of synchronization on dynamic grouping, and of context on visual size perception will be described. Evidence for the relevance of coordinating interactions to cognitive style and to cognitive disorganization in psychosis will be outlined.

Figure-Ground Segmentation by Integration of Multiple Cues

Jan-Olof Eklundh, Mårten Björkman and Eric Hayman

Computational Vision and Active Perception Laboratory (CVAP)
Dept. of Numerical Analysis and Computing Science
Royal Institute of Technology (KTH), SE 100 44 Stockholm, Sweden
{joe, celle, hayman}@nada.kth.se

Humans looking around in the world can, seemingly without effort, segment out and distinguish different objects in the world. The corresponding capability has largely eluded the efforts of researchers in computer vision. The problem is of course not well-defined unless additional assumptions are made. If we ask ourselves what objects we see around us we realize that such a question has little meaning and is too imprecise to answer. We need at least a model of the visual observer and the tasks this observer is engaged in to specify what these objects could be. They are not given by the visual scene alone.

The processes of perceiving objects in the world and segmenting images of them depend on each other and figure-ground segmentation is generally not feasible solely bottom-up. Whatever the processes are they should be possible to bootstrap in some way and a question is what such mechanisms could be. Work in perceptual grouping and attention address some such aspects. Here we'll discuss the use of 3D cues for figure-ground separation. If something stands out in 3D, then it forms a separate piece of materia and as such it is more than something that just stands out visually as, say, a set of contours, surface markings or colored patches. Such visual patterns may indicate objects or groups of objects in a multitude of ways. Unless we know more about the scene it is difficult to say if they define any relevant objects. On the other hand, even without such knowledge, we can identify a 3D chunk as "something", which then can be ascribed visually observable 3D properties, such as position, location and motion, but also object intrinsic properties such as shape, color and maybe surface and material characteristics.

One thing the 3D cues tell us is the geometric relation between the observer and the object. The identity of a (not necessarily recognized or labeled) object can also be maintained over time by the appearance of the object, i.e. properties such as shape and color that can be obtained from the 2D images. This strongly suggests that a system for robust figure-ground segmentation in a dynamically changing world should rely on mutiple cues in 3D and 2D and that the 3D cues play a specific role in the bootstrapping. This forms a main theme of the talk. In it we will discuss some of the underlying issues and illustrate them with examples from our own work.

We will discuss figure-ground segmentation based on stereo and motion cues together with monocular cues from e.g. texture. We will also discuss the combination of purely monocular cues from motion, color and contrast. We will consider

several different integration techniques. One is a probabilistic approach where the likelihood of observing the data given a model of each layer is computed followed by a classification of each pixel using Bayes' rule. A second scheme is a voting method, the key difference being that each cue makes an independent decision regarding membership before these decisions are combined using a weghted sum. The advantage of voting in data fusion is that measurements drawn from very different spaces can easily be combined. With probabilistic methods more care must be taken in designing the model of each so that the different cues combine in the desired manner. However, in that also lies there strenght since it requires an explicit design of the model and specification of what parameters are used and what assumptions are made.

In the binocular case we show that even coarse estimates of relative depth provide information that strongly facilitates the computation of motion and through feedback also depth.

There are many algorithms available for computing the specific cues. Some of these require iterative solutions and are therefore not always suited for use in a full-fledged integrated system working in a real dynamic world, since they cause serious delays. We will therefore go through a number of different algorithms from a complexity and precision point of view and present results both on simulated and real data.

A final aspect concerns how models for some cues can be learnt and subsequently be adpated online. For instance, this applies to the case when 3D cues indicate an object for which we can learn some appearance properties, e.g. a color model over time. We'll show some results in this direction. One of the main motivations for this work is in fact to give support to high level visual processes, such as recogntion and categorization in realistic and natural environments.

References

1. M. Björkman, *Real-time motion and stereo cues for an active visual observer*, Doctoral disseration, TRITA-NA-0213, KTH, Stockholm, June 2002.
2. M. Björkman and J-O. Eklundh, Real-time epipolar geometry estimation of binocular stereo headsm, *IEEE Trans. PAMI*, Vol 24:3, 425-432, March 2002.
3. E. Hayman and J-O. Eklundh, Probabilistic and voting approaches to cue integration for figure-ground segmentation, In *Proc. ECCV*, Springer LNCS, Vol 2352, 469-486, May 2002.

Optic flow

Detection of First-order Elementary Components in Noisy Optic Flow Fields Through Context Sensitive Recurrent Filters

Silvio P. Sabatini, Fabio Solari, and Giacomo M. Bisio

Dept. of Biophysical and Electronic Engineering, University of Genoa
Via all'Opera Pia, 11a - 16145 Genova, Italy
{silvio, fabio, bisio}@dibe.unige.it
http://www.pspc.dibe.unige.it

Abstract. Measured optic flow fields are always somewhat erroneous and/or ambiguous. First, we cannot compute the actual spatial or temporal derivatives, but only their estimates, which are corrupted by image noise. Second, optic flow is intrinsically an image-based measurement of the relative motion between the observer and the environment, but we are interested in estimating the actual motion field. However, real-world motion field patterns contain intrinsic statistic properties that allow to define Gestalts as groups of pixels sharing the same motion property. By checking the presence of such Gestalts in optic flow fields we can make their interpretation more confident. We propose an optimal recurrent filter capable of evidencing motion Gestalts corresponding to 1st-order spatial derivatives or elementary flow components (EFCs). A Gestalt emerges from a noisy flow as a solution of an iterative process of spatially interacting nodes that correlates the statistics of the visual context with that of a structural model of the Gestalt.

1 Local motion Gestalts

Velocity gradients provide important cues about the 3-D layout of the visual scene. Formally, they can be described as *linear deformations* by a 2×2 velocity gradient tensor

$$\mathbf{T} = \begin{bmatrix} T_{11}\ T_{12} \\ T_{21}\ T_{22} \end{bmatrix} = \begin{bmatrix} \partial v_x/\partial x\ \partial v_x/\partial y \\ \partial v_y/\partial x\ \partial v_y/\partial y \end{bmatrix} . \tag{1}$$

Hence, if $\boldsymbol{x} = (x, y)$ is a point in a spatial image domain, the linear properties of a motion field $\boldsymbol{v}(x, y) = (v_x, v_y)$ around the point $\boldsymbol{x}_0 = (x_0, y_0)$ can be characterized by a Taylor expansion, truncated at the first order:

$$\boldsymbol{v} = \bar{\boldsymbol{v}} + \bar{\mathbf{T}}\boldsymbol{x} \tag{2}$$

where $\bar{\boldsymbol{v}} = \boldsymbol{v}(x_0, y_0) = (\bar{v}_x, \bar{v}_y)$ and $\bar{\mathbf{T}} = \mathbf{T}|_{\boldsymbol{x}_0}$. By breaking down the tensor in its dyadic components, the motion field can be locally described through 2-D maps representing *cardinal* EFCs:

$$\boldsymbol{v} = \boldsymbol{\alpha}^x \bar{v}_x + \boldsymbol{\alpha}^y \bar{v}_y + \boldsymbol{d}_x^x \left.\frac{\partial v_x}{\partial x}\right|_{\boldsymbol{x}_0} + \boldsymbol{d}_y^x \left.\frac{\partial v_x}{\partial y}\right|_{\boldsymbol{x}_0} + \boldsymbol{d}_x^y \left.\frac{\partial v_y}{\partial x}\right|_{\boldsymbol{x}_0} + \boldsymbol{d}_y^y \left.\frac{\partial v_y}{\partial y}\right|_{\boldsymbol{x}_0} \tag{3}$$

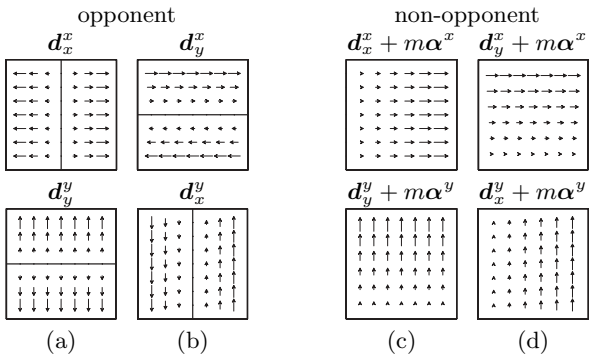


Fig. 1. Basic gradient type Gestalts considered. In stretching-type components (a,c) velocity varies *along* the direction of motion; in shearing-type components (b,d) velocity gradient is oriented *perpendicularly* to the direction of motion. Non-opponent patterns are obtained from the opponent ones by a linear combination of pure tranlations and cardinal deformations: $\boldsymbol{d}_j^i + m\boldsymbol{\alpha}^i$, where m is a proper positive scalar constant.

where $\boldsymbol{\alpha}^x : (x, y) \mapsto (1, 0)$, $\boldsymbol{\alpha}^y : (x, y) \mapsto (0, 1)$ are pure translations and $\boldsymbol{d}_x^x : (x, y) \mapsto (x, 0)$, $\boldsymbol{d}_y^x : (x, y) \mapsto (y, 0)$, $\boldsymbol{d}_x^y : (x, y) \mapsto (0, x)$, $\boldsymbol{d}_y^y : (x, y) \mapsto (0, y)$ represent cardinal deformations, basis of the linear deformation space.

It is worthy to note that the components of pure translations could be incorporated in the corresponding deformation components, thus obtaining generalized deformation components in which motion boundaries are shifted or totally absent. Although this does not affect the significance of the Taylor expansion in Eq. 3, the so-modified elementary components, present very different structural properties. Since a template-based approach cannot be used to extract single components, but only to perform pattern matching operations, the linear decomposition of the motion field has significance only for the definition of a proper representation space. Specific templates would be designed to optimally sample that representation space. In this work, we consider two different classes of deformation templates (opponent and non-opponent), each characterized by two gradient types (stretching and shearing), see Fig. 1. Due to their ability to detect the presence and the orientation of velocity gradients and kinetic boundaries, such cardinal EFCs and proper combinations of them resemble the characteristics of the cell in the Middle Temporal visual area (MT) [1] [2]. It is straightforward to derive that these MT-like components are well suited to provide the building blocks for the more complex receptive field properties encountered in the Medial Superior Temporal visual area (MST) [3] [4]:

$$\boldsymbol{v} = \boldsymbol{\alpha}^x \bar{v}_x + \boldsymbol{\alpha}^y \bar{v}_y + \frac{1}{2}(\boldsymbol{d}_x^x + \boldsymbol{d}_y^y)E + \frac{1}{2}(\boldsymbol{d}_x^x - \boldsymbol{d}_x^y)\omega + \frac{1}{2}(\boldsymbol{d}_x^x - \boldsymbol{d}_y^y)S_1 + \frac{1}{2}(\boldsymbol{d}_y^x + \boldsymbol{d}_x^y)S_2$$

where $E = (\bar{T}_{11} + \bar{T}_{22})/2$, $\omega = (\bar{T}_{12} - \bar{T}_{21})/2$, $S_1 = (\bar{T}_{11} - \bar{T}_{22})/2$, $S_2 = (\bar{T}_{12} + \bar{T}_{21})/2$ are the divergence, the curl and the two components of shear

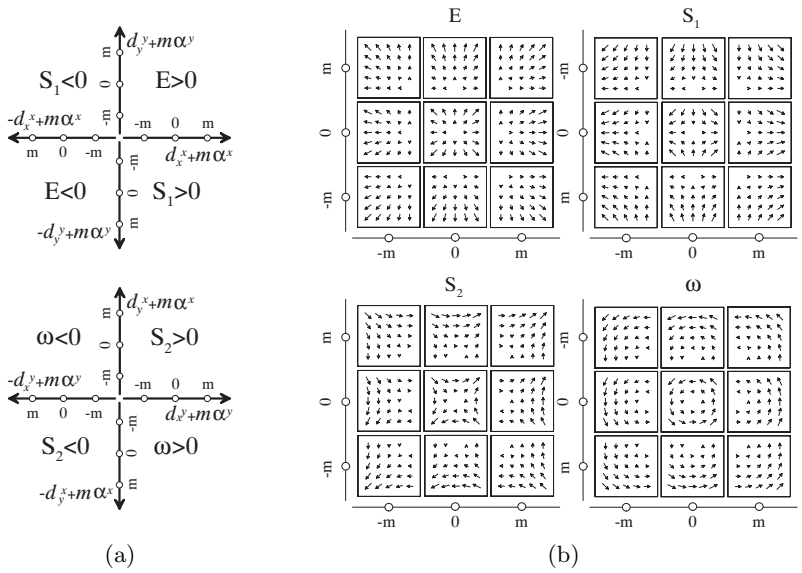


Fig. 2. (a) Two deformation subspaces obtained by the set of cardinal EFCs with different values of the parameter m. The quadrants of each subspace characterize an elementary deformation, as evidenced in (b) for expansion ($E > 0$), horizontal positive shear ($S_1 > 0$), oblique positive shear (S_2), and counterclockwise rotation ($\omega > 0$).

deformation, respectively (cf. [5]). These mixed EFCs constitute, together with the pure translations, an equivalent representation basis for the linear properties of the velocity field (see Fig. 2). Yet, they are rather complex since not only the speed, but also the direction of feature motion varies as a function of spatial position. Rigid body motion often generates simpler flow fields characterized by unidirectional patterns, as the cardinal EFCs considered in this study.

2 The context sensitive filter

The problem of evidencing the presence of a certain complex feature in the optic flow on the basis of both local and contextual information, can be posed as an adaptive filtering problem (estimation), where local information act as the input *measurements* and the context acts as the *reference signal*, e.g., representing a specific motion Gestalt. In the following, we propose a solution in the form of a generalized Kalman filter (KF) [6]. Due to its recurrent formulation, KF appears particularly promising to design *context-sensitive filters* (CSFs) based on recurrent cortical-like interconnection architectures.

Let us assume the optic flow $\tilde{\boldsymbol{v}}(i,j)$ as the corrupted measure of the actual velocity field $\boldsymbol{v}(i,j)$. The difference between these two variables can be represented

as a constant noise term $\boldsymbol{\varepsilon}(i,j)$:

$$\tilde{\boldsymbol{v}} = \boldsymbol{v} + \boldsymbol{\varepsilon}\ . \tag{4}$$

Due to the intrinsic noise of the nervous system, the neural representation of the optic flow $\mathbf{v}(i,j)[k]$ can be expressed by a *measurement equation*:

$$\mathbf{v}[k] = \tilde{\boldsymbol{v}} + \boldsymbol{n}_1[k] = \boldsymbol{v} + \boldsymbol{\varepsilon} + \boldsymbol{n}_1[k] \tag{5}$$

where $\boldsymbol{n}_1$ represents the uncertainty associated with a neuron's response. The Gestalt is formalized through a *process equation*:

$$\boldsymbol{v}[k] = \boldsymbol{\Phi}\boldsymbol{v}[k-1] + \boldsymbol{n}_2[k-1] + \boldsymbol{s} \tag{6}$$

with $\lim_{k\to\infty} \boldsymbol{v}[k] = \boldsymbol{v}$ if $\boldsymbol{n}_2 = 0$. The state transition matrix $\boldsymbol{\Phi}$ is *de facto* a spatial interconnection matrix that implements a specific Gestalt rule (i.e., a specific EFC); $\boldsymbol{s}$ is a constant driving input; $\boldsymbol{n}_2$ represents the process uncertainty. The space spanned by the observations $\mathbf{v}[1]$, $\mathbf{v}[2]$,..., $\mathbf{v}[k-1]$ is denoted by $\boldsymbol{\mathcal{V}}_{k-1}$ and represents the internal noisy representation of the optic flow. We assume that both $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ are independent, zero-mean and normally distributed: $\boldsymbol{n}_1[k] = N(0, \boldsymbol{\Lambda}_1)$ and $\boldsymbol{n}_2[k] = N(0, \boldsymbol{\Lambda}_2)$. The index k takes explicitly into account the time necessary for spatial recurrence. More precisely, $\boldsymbol{\Phi}$ models space-invariant nearest-neighbor interactions within a finite region Ω in the (i,j) plane that is bounded by a piece-wise smooth contour. Interactions occur, separately for each component of the velocity vectors (v_x, v_y), through anisotropic interconnection schemes:

$$\begin{aligned} v_{x/y}(i,j)[k] &= w_N^{x/y} v_{x/y}(i,j-1)[k-1] + w_S^{x/y} v_{x/y}(i,j+1)[k-1] + s_{x/y}(i,j) \\ &+ w_W^{x/y} v_{x/y}(i-1,j)[k-1] + w_E^{x/y} v_{x/y}(i+1,j)[k-1] + n_1^{x/y}(i,j)[k-1] \end{aligned}$$

where (s_x, s_y) is a steady additional control input, which models the boundary conditions. The process equation has a *structuring effect* constrained by the boundary conditions that yields to structural equilibrium configurations, characterized by specific first-order EFCs. The resulting pattern depends on the anisotropy of the interaction scheme and on the boundary conditions. By example, considering, for the sake of simplicity, a rectangular domain $\Omega = [-L, L] \times [-L, L]$, the cardinal EFC $\boldsymbol{d}_x^x$ can be obtained through:

$$\begin{array}{ll} w_N^x = w_S^x = 0 & w_N^y = w_S^y = 0 \\ w_W^x = w_E^x = 0.5 & w_W^y = w_E^y = 0 \end{array} \quad s_x(i,j) = \begin{cases} -\lambda \text{ if } i = -L \\ \lambda \quad \text{if } i = L \\ 0 \text{ otherwise} \end{cases} \quad s_y(i,j) = 0$$

where the boundary value λ controls the gradient slope. In a similar way we can obtain the other components.

Given Eqs. (5) and (6), we may write the optimal filter for optic flow Gestalts. The filter allows to detect, in noisy flows, intrinsic correlations, as those related to EFCs, by checking, through spatial recurrent interactions, that the spatial context of the observed velocities conform to the Gestalt rules, embedded in $\boldsymbol{\Phi}$.

To understand how the CSF works, we define the *a priori* state estimate at step k given knowledge of the process at step $k-1$, $\hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k-1}]$, and the *a posteriori* state estimate at step k given the measurement at the step k, $\hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k}]$. The aim of the CSF is to compute an *a posteriori* estimate by using an *a priori* estimate and a weighted difference between the current and the predicted measurement:

$$\hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_k] = \hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k-1}] + \boldsymbol{G}[k]\ (\mathbf{v}[\mathbf{k}] - \hat{\mathbf{v}}[\mathbf{k}|\boldsymbol{\mathcal{V}}_{\mathbf{k-1}}]) \tag{7}$$

The difference term in Eq. (7) is the *innovation* $\boldsymbol{\alpha}[k]$ that takes into account the discrepancy between the current measurement $\mathbf{v}[\mathbf{k}]$ and the predicted measurement $\hat{\mathbf{v}}[\mathbf{k}|\boldsymbol{\mathcal{V}}_{\mathbf{k-1}}]$. The matrix $\boldsymbol{G}[k]$ is the Kalman gain that minimizes the *a posteriori* error covariance:

$$\boldsymbol{K}[k] = E\left\{(\boldsymbol{v}[k] - \hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_k])(\boldsymbol{v}[k] - \hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_k])^T\right\}\ . \tag{8}$$

Eqs. 7 and 8 represent the mean and covariance expressions of the CSF output.

The covariance matrix $\boldsymbol{K}[k]$ provides us only information about the properties of convergence of the KF and not whether it converges to the correct values. Hence, we have to check the consistency between the innovation and the model (i.e., between observed and predicted values) in statistical terms. A measure of the reliability of the KF output is the Normalized Innovation Squared (NIS):

$$NIS_k = \boldsymbol{\alpha}^T[k]\ \boldsymbol{\Sigma}^{-1}[k]\ \boldsymbol{\alpha}[k] \tag{9}$$

where $\boldsymbol{\Sigma}$ is the covariance of the innovation. It is possible to exploit Eq. (9) to detect if the current observations are an instance of the model embedded in the KF [7].

3 Results

Fig. 3 shows the responses of the CSF in the deformation subspaces for two different input flows. Twentyfour EFC models have been used to span the deformation subspaces shown in Fig. 2a. The grey level in the CSF output maps represents the probability of a given Gestalt according to the NIS criterium: lightest grey indicates the most problable Gestalt. Besides Gestalt detection, context information reduces the uncertainty on the measured velocities, as evidenced, for the circled vectors, by the Gaussian densities, plotted over the space of image velocity.

4 Conclusions

Given motion information represented by an optic flow field, we specified a CSF to recognize if a group of velocity vectors belong to a specific pattern, on the basis of their relationships in a spatial neighborhood. Casting the problem as a KF, the detection occurs through a spatial recurrent filter that checks the consistency between the spatial structural properties of the input flow field pattern

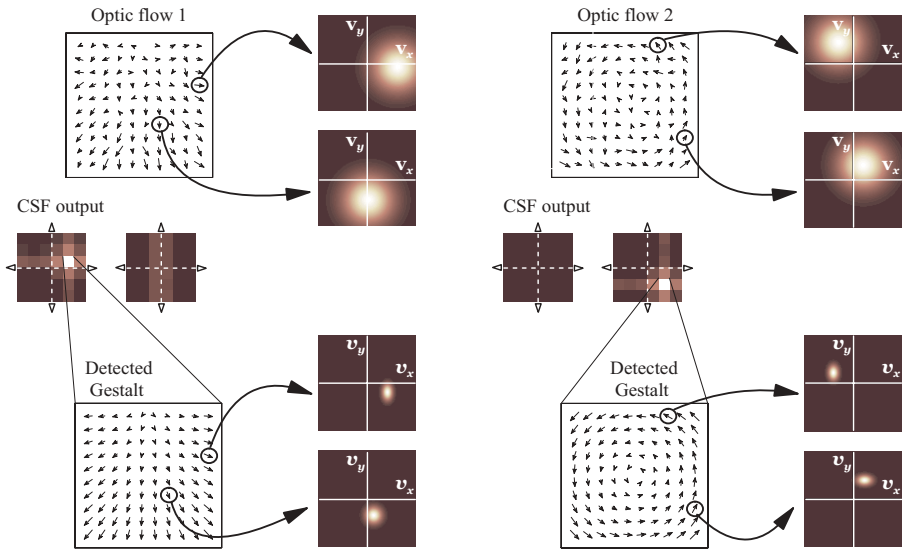


Fig. 3. Example of Gestalt detection in noisy flows.

and a structural rule expressed by the process equation of the KF. The CSF behaves as a template model. Yet, its specificity lies in the fact that the template character is not built by highly specific feed-forward connections, but emerges by stereotyped recurrent interactions (cf. the process equation). Furthermore, the approach can be straightforwardly extended to consider adaptive cross-modal templates (e.g, motion and stereo). By proper specification of the matrix $\mathbf{\Phi}$, the process equation can, indeed, potentially model any type of multimodal spatio-temporal relationships (i.e., multimodal spatio-temporal context).

References

1. V.L. Marcar, D.K. Xiao, S.E. Raiguel, H. Maes, and G.A. Orban. Processing of kinetically defined boundaries in the cortical motion area MT of the macaque monkey. *J. Neurophysiol.*, 74(3):1258–1270, 1995.
2. S. Treue and Andersen R.A. Neural responses to velocity gradients in macaque cortical area MT. *Visual Neuroscience*, 13:797–804, 1996.
3. C.J. Duffy and R.H. Wurtz. Response of monkey MST neurons to optic flow stimuli with shifted centers of motion. *J. Neuroscience*, 15:5192–5208, 1995.
4. M. Lappe, F. Bremmer, M. Pekel, A. Thiele, and K.P. Hoffmann. Optic flow processing in monkey STS: A theoretical and experimental approach. *J. Neuroscience*, 16:6265–6285, 1996.
5. J.J. Koenderink. Optic flow. *Vision Res.*, 26(1):161–179, 1986.
6. S. Haykin. *Adaptive Filter Theory.* Prentice-Hall International Editions, 1991.
7. Y. Bar-Shalom and X.R. Li. *Estimation and Tracking, Principles, Techniques, and Software.* Artech House, 1993.

Cortical mechanisms of processing visual flow – Insights from the Pinna-Brelstaff illusion

Pierre Bayerl and Heiko Neumann

Department of Neural Information Processing, University of Ulm, Germany,
{pierre, hneumann}@neuro.informatik.uni-ulm.de

1 Introduction

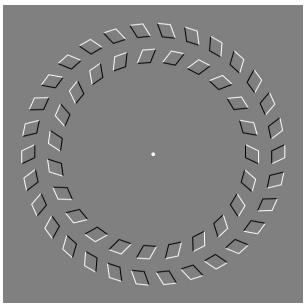


Fig. 1. The illusion of relative motion introduced by Pinna and Brelstaff [1]

It remains an open question how different cortical areas interact to accomplish the robust analysis of moving visual patterns. In order to gain insights of the neural mechanisms underlying the cortical processing of large-field motion patterns, we investigate a relative motion illusion presented by Pinna and Brelstaff [1]. The stimulus pattern consists of circularly arranged tiles each bounded by light and dark lines (Fig.1, left). A forward and backward moving human observer induces a strong illusory motion of clockwise and counter-clockwise rotation of the inner and outer ring while fixating the center of the circular arrangements of tiles. The contrast arrangement along the boundary of individual tiles as well as between the tiles and the peripheral location of the items is important to generate the illusion. We claim that an investigation of the input-output relation between stimulus and (illusory) percept reveals key principles of the neural processing of flow patterns in the dorsal pathway.

We developed a model of recurrent interaction of areas V1, MT, and MSTd along the dorsal cortical pathway utilizing a space-variant mapping of flow patterns [2]. The model predicts the perception of relative motion for the Pinna-Brelstaff pattern and new variants of it. In this paper these predictions were psychophysically investigated in order to assess the strength of relative motion in a parametric fashion.

2 Model and computational results

Motion information is processed primarily along cortical pathways which involve areas V1, V2, MT, and MSTd, respectively. Our model [2] is based on a space-variant representation of V1 as proposed by Schwartz [3]. Motion information is integrated along the V1-MT-MSTd feed-forward pathway utilizing direction selective cells of increasing spatial size (1:11:30 ratio). Directional inhibition

modeled at the stage of MT [4] explains why certain configurations of the input pattern yield no illusory effect. Most important to our model is that salient patterns are detected with less spatial accuracy in higher areas. The resulting activities are fed back to disambiguate information at higher spatial resolution provided in earlier areas. In the investigated motion illusion, modulatory MSTd-MT feedback achieves necessary disambiguation of initial unspecific optic flow estimates. This leads to segregated opponent motions along circular directions when perceptual splitting occurs, while homogeneous motion fields are detected when no splitting is observed. Some results of computational simulations are sketched in Fig. 2. Concerning different contrast configurations of the original stimulus, our model simulations are consistent with the findings of Pinna and Brelstaff. The difference between input patterns with one ring and patterns with two rings is that directional decomposition only occurs for illusory stimuli consisting of two rings. This decomposition punctuates the rotational part of illusory motion and enhances perceptual splitting of both rings. The existence of a mechanism which segregates adjacent flow regions and disambiguates flow estimations is stressed by the following experiments.

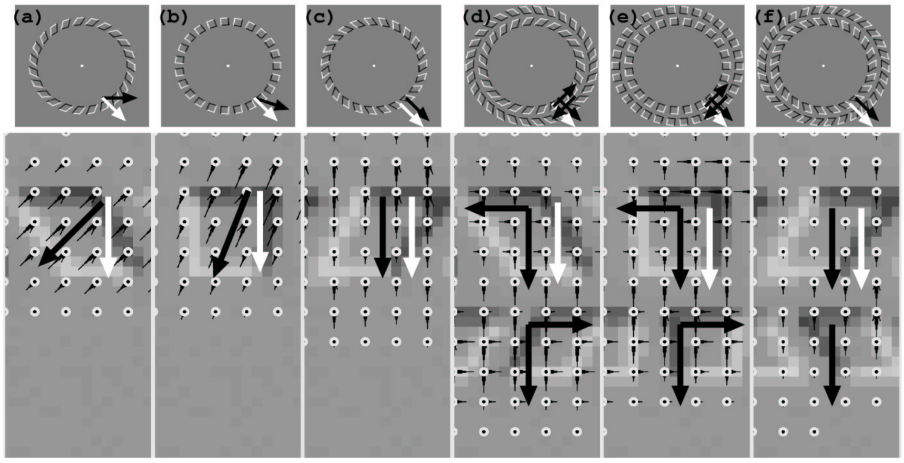


Fig. 2. Simulation results for different contrast configurations of the original stimulus. The dots and lines encode activities of model MT cells sensitive to the indicated direction (lines) at the corresponding location (dots). The background represents a cutout of the log-polar mapped input stimulus (circular directions are plotted along the abscissa, radial directions along the ordinate). Model MT activities are the result of several iterations of feedback processing and therefore are already completely disambiguated. Dark arrows indicate the mean directions of detected motion components, light arrows the direction of true motion. **(a-c):** Stimulus configurations with one ring of tiles, illusory patterns (a,b) and non-illusory pattern (c). **(d-f):** Stimulus configurations with two rings of tiles: Note that directional decomposition (perceptual splitting) occurs for the illusory patterns (d,e) and not for the non-illusory pattern (f).

3 Psychophysical experiments and results

We developed an experimental setup to test variants of the Pinna-Brelstaff illusion that were predicted by our model. This allows to quantify the relative speed of various motion patterns in a parametric fashion. These results can be used again to verify the model predictions by imposing the respective pattern to the neural computational model.

General Stimuli Configuration: The stimuli consist of one or two rings containing circularly arranged tiles of a certain type. Beside the original tiles we investigated patterns composed of patches of oriented Gabor wavelets[1], which allow to parameterize spatial frequencies and orientations of the stimulus. A novel variation is the additive combination of two different Gabor patterns to induce two different motion cues at the same location. The displays of looming ring patterns are generated using real-time computer graphics techniques (OpenGL). The speed v of true radial motion is held constant. All stimulus parameters like speed or wavelengths are specified in pixel, one pixel corresponds approximately to 0.026 degrees at a viewing distance of 60 cm.

Task/Procedure: In a nulling task an observer is asked to parametrize real spiral motions to counteract the illusion perceived for the inner ring that is induced by the looming pattern. In order to get accurate results in an acceptable amount of time the Best PEST method [5] is applied to detect the threshold of nulling the illusory effect. The rotational correction is applied anti-symmetrical on both rings. This correction does not affect the inner and the outer ring equally: some configurations exist for which the outer ring still induces an illusion of relative motion after the rotational components of the inner ring have been eliminated. The results however show that the amount of correction applied for the inner ring correlates with the strength of the illusion reported by Pinna and Brelstaff and the predictions of our model.

3.1 Experiment 1: contrast orientation of the original tiles

In the first experiment we acquire psychophysical data, which can directly be compared to our computational results concerning different contrast configurations of the original illusion presenting either both rings or only the inner ring. Pinna and Brelstaff found that certain contrast configurations yield a stronger illusory effect than others, but it remains unclear if this effect is influenced by spatial interactions between both rings or not. In particular, we want to know whether the illusory effect is influenced by directional repulsion caused by a motion contrast between both rings. Stimuli are tested for three contrast configurations as presented in [1]. We varied the shearing angle α of the tiles ($\alpha \in \{-40, 0, 40\}$, see Fig. 3) either with both rings or the inner ring only.

Results: The results illustrated in Fig. 3 (right) show the amount of rotational correction for different stimulus parameters. The results for two rings are

[1] Only recently we got notice that earlier this year Mike Morgan utilized a similar variant of such stimulus for demonstration purposes.

qualitatively consistent with the findings of Pinna and Brelstaff and with our computational results (Fig. 3, left). No significant difference can be observed for different numbers of rings. We conclude that for the investigated stimuli there is no significant interaction between both rings that influences the strength of the final percept. If the final perceived motion is interpreted as a population vector represented by cells within a ring, the psychophysical findings for the single-ring stimuli are also consistent with our model simulations. The observation reported by Pinna and Brelstaff that apparent rotations of single-ring stimuli appear to be much weaker compared to patterns with two rings could be explained by the optical flow decomposition performed by our model. This decomposition stresses the existence of illusory rotational motion components.

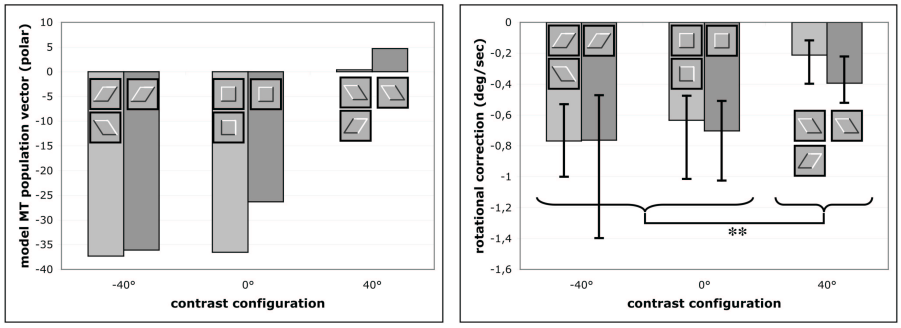


Fig. 3. Results for different contrast configurations with one and two rings of the original stimulus tiles. Left: model predictions for the direction of the MT population vector. (inner ring) Right: median plot of psychophysical data (Exp. 1, mean±min/max, N=8 trials). The illusory effect for the first two configurations (shear = -40^o and 0^o) is significant stronger (**=p≤.01, U-Test) than for the third configuration (shear = 40^o).

3.2 Experiment 2: oriented Gabor patches

In order to investigate the role of contrast orientation in more detail we propose a stimulus setup using oriented Gabor patches. Gabor wavelets have the advantage to induce a motion cue for a specific scale (λ) and direction (α, $\alpha = 0$ means radial orientation). If the aperture problem would explain the illusion as proposed by Pinna and Brelstaff, the strength of the stimulus should be proportional to $\sin(\alpha)\cos(\alpha)$ and therefore maximized for $\alpha = 45^o$ with a local symmetry around $\alpha = 45^o$. Effects of interaction between both rings are re-examined because directional repulsion or motion contrast enhancement may only occur for small angular differences. Stimuli are tested for 8 contrast orientations ($\alpha \in \{\pm 67.5^o, \pm 45^o, \pm 22.5^o, \pm 11.25^o\}$) either using both rings or the inner ring only.

Results: Data shown in Fig. 4 (left) demonstrates that the responses are not distributed symmetrical around $\alpha = 45^o$. Also stimuli with two rings show enhanced illusory effects for small values of α compared to the single-ring stimuli. Therefore some mechanism seems to generate a directional repulsion within an area covering both rings as well as within the rings. We propose that the repulsion

is initiated in MSTd and that modulatory feedback separates the model MT responses to form the final percept. Due to increasing receptive field sizes and therefore decreasing spatial accuracy, flow information of both rings at the stage of MSTd is likely to be handled as motion transparency. An alternative to the directional decomposition performed by our model MSTd would be a mechanism of directional repulsion as proposed by Kim and Wilson for their model of motion transparency [6].

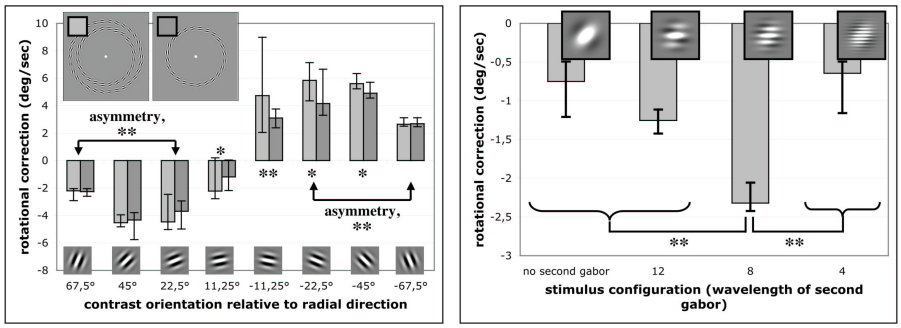
3.3 Experiment 3: compound Gabor patches

Our model predics that the detection of the true radial flow might enhance the repulsion of illusory flow components. To test if the illusion is enhanced by such a directional interaction we generate stimulus tiles, which consist of compound Gabor patches (additive combination of two Gabor wavelets) inducing flow information for spiral and radial directions (*compound stimulus*): A low-frequency Gabor patch generates clockwise and counter-clockwise spiral motion cues for the inner and outer ring, respectively ($\alpha_1 = 45^o, \lambda = 38$). An overlaid Gabor patch with higher frequency induces radial flow information ($\alpha_2 = 0^o$). Different wavelengths for the radial oriented Gabor are investigated: $\lambda \in \{12, 8, 4\}$. We also tested a stimulus configuration with tiles containing a single Gabor ($\alpha_1 = 45^o, \lambda = 38$) without overlay (*uniform stimulus*).

Results: The results (Fig. 4, right) provide information of specific interactions of cells tuned to different directions and different spatial frequencies. For the compound stimulus with $\lambda = 8$ the enhancement of the effect compared to the uniform stimulus is very significant. Also compound stimuli with lower frequencies show an increased illusory effect. Only for the highest frequency ($\lambda = 4$) the enhancement collapses. This might be caused by the fact that the visual system is unable to detect such high frequencies in the periphery. These findings stress the role of directional repulsion between different directions of motion induced by patterns of different scales. Like in experiment 2 this effect can be explained by a mechanism of directional repulsion. An alternative, but rather speculative, interpretation is the following: the illusory percept for compound stimuli is the result of mechanisms combining form cues with motion cues like those observed for the barberpole illusion[7]. Most essential for this explanation is that the high-frequency Gabor patch cannot be accurately located due to its eccentricity. The radial wave fronts induced by this Gabor patch act as (static) circular boundaries. The perceived illusory rotation between these circular boundaries could then be explained with mechanisms including form information from the form pathway like those in the model of Viswanathan [7].

4 Summation and conclusion

Our neural computational model [2] provides evidence that feedback from the higher-order motion area MST is essential for generating unambiguous patterns of large-field motion. Our investigations also led to a novel interpretation of the



(**=p≤.01, *=p≤.05, U-Test)

Fig. 4. Left (Exp. 2, mean±min/max, N=9 trials): Strength of the illusion for different orientations (α) of Gabor patches. The stimuli consist of either one or two rings of patches. For some configurations the amount of rotational correction is significant higher for two rings. Most important is that the responses are not symetrically distributed around $\alpha = \pm 45^o$ as predicted by the simple normal flow model[1]. **Right** (Exp 3, mean±min/max, N=8 trials): Strength of the illusion for different configurations of *compound stimuli* compared with an *uniform stimulus* configuration. To generate compound stimuli a high frequency, radial oriented Gabor patch ($\lambda = 4, 8, 12$, $\alpha = 0^o$) is added to a low frequency, diagonal oriented Gabor patch ($\lambda = 38$, $\alpha = 45^o$). The uniform stimulus consists of the diagonal oriented Gabor patch without overlay. For $\lambda = 8$ the illusory effect of compound stimuli is almost doubled compared to the uniform stimulus and also significant stronger than all other configurations.

Pinna-Brelstaff illusion as one of motion transparency [4, 6] in which the same mechanisms are involved to generate the observed perceptual segregations.

Our experimental investigations using the original stimulus tiles reproduce the results obtained by Pinna and Brelstaff. With our nulling technique we now quantified the strength of the illusion in relation to other variants of the input pattern. Experiments concerning the role of the component spatial frequencies and their orientations reveal evidence for specific interactions between cells tuned to different motion directions and different spatial frequencies. In particular we found that the investigated illusions cannot be ascribed solely to the aperture effect and that additional mechanisms are needed like those presented in [2].

References

1. Pinna, B., Brelstaff, G.J. Vision Research, **40** (2000)
2. Bayerl, P.A.J., Neumann, H. BMCV (2002), in print.
3. Schwartz, E.L. In A. Peters and K. Rocklund, editors, Cerebral Cortex, **10** (1994)
4. Qian, N., Andersen, R.A., Adelson, E.H. The Journal of Neuroscience, **14** (1994)
5. Lieberman, H.R., Pentland, A.P. Behavior Res. Methods & Instr., **14** (1982)
6. Kim, J., Wilson, H.R.: Vision Research, **36** (1996)
7. Grossberg, S., Mingolla, E., Viswanathan, L. Vision Research, **41** (2001)

Integration of landmark information and optic flow in humans

Sabine Gillner and Yu Jin

University of Tübingen, Institute for Cognitive Neuroscience, Morgenstelle 28, 72076 Tübingen, Germany,
sabine.gillner@uni-tuebingen.de,
WWW home page: http://www.uni-tuebingen.de/cog

Abstract. We investigated the ability of humans to combine landmark information and optic flow. 20 subjects (11 male, 9 female) participated in a desktop virtual reality experiment consisting of six stages. In five experiments only optic flow was available to the subjects. In two of this five experiments the subjects received feedback by presenting a birdseye view of their journey. In the last experiment a landmark was introduced, which was replaced in some of the trails during the testphase, not informing the subjects about that. The landmark has had an important influence on the computing of the homevector, but also the optic flow influenced the result.

1 Introduction

Path integration, one of several possible spatial navigation mechanisms, is the process of determining one own's position on the basis of egomotion. Loomis et al. [3] investigated this ability in blind and blind–folded people and could show that human subjects tend to show a systematic error in that way that shorter distances are overestimated and longer distances are underestimated. While these results were based on non–visual information, optic flow could also contribute to the process of path integration. It has been shown in several studies that humans are very well capable of estimating their heading direction from very short presentations of a flow field [7] but it is also known that it is much more difficult to estimate a longer trajectory from this information [1]. Recently, Riecke et al. [4] replicated the Loomis et al. study in a visual task. They used a 180^o–projection screen and did find a much smaller error than in the former experiments. They could also suppress the compression–to–the mean error in certain experimental conditions. But it is not clear how relevant optic flow information in a navigation task is, if additional information is available. For human spatial cognition, landmarks play a dominant role to lead a navigator to his goal. Landmarks could be used in a different manner: they provide local position information or could be used as a course maintaining aid. In this latter sense we investigated the role of landmarks in this paper. We were interested in the question if – and how this landmark information is combined with the information from optic flow.

a) b)

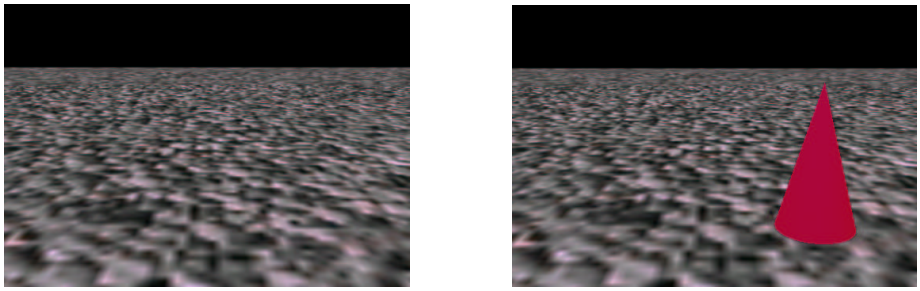


Fig. 1. Screenshots of the experiment a) without and b) with a landmark. These views have been taken from the starting point. During the passage of the landmark it was always to the right side of the subjects.

2 Methods

The experiment was run on a personal computer (Linux – 700 MHz PC, NVidea GForce2). 20 volunteered, paid subjects (11 female, 9 male) participated in this experiment. They sat in front of the computer screen in a comfortable distance without any head– and chinrest. The screen spanned therefore 30–40 deg of the horizontal field–of–view of the subjects.

OpenPerformer and C++ were used for the programming of the experimental environment. It consists of a textured ground floor only. In some of the trials a landmark was introduced (see figure 1). Egomotion was simulated after pressing the corresponding cursor buttons (i.e. left arrow $\rightarrow$ turning to the left), subjects could either turn (10 deg $\sec^{-1}$) or move straight forward or backward (1 m $\sec^{-1}$)[1]. As the experimental procedure we used the triangle completion paradigm (see figure 2) where subjects were guided passively along two legs and the including angle of a triangle. Then they had to "walk" back to the starting point by pressing the cursor buttons indicating which homevector they had built. The experiment consisted of six stages: five stages without landmark, the

[1] In our setup the notation "Meter" is arbitrary, in a strikt sense we should use the term "Unit".

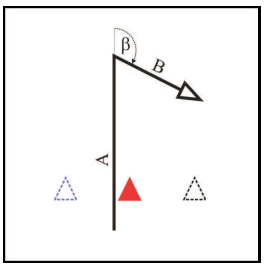


Fig. 2. Setup of the landmark trials. Subjects were led two legs of the triangle (Leg A and B in the figure). In some trials a red cone served as a landmark, placed in the vicinity of the starting point ($x = 0.5m, y = 1m$). After the passage of this landmark, it was either shifted 5m to the right or to the left (dashed triangles), or it remained at the same position. The subjects did not see this transposition. We varied the length of leg A between 2,6 and 10 m. Leg B was always 2m. The turning angle β has had one of the following values: $-120^o, -90^o, -60^o, 60^o, 90^0, 120^o$
.

	60^o	90^o	120^o	-60^o	-90^o	-120^o
2 m	←	→	o	o	←	→
6 m	*o*	←	→	←	→	o
10 m	→	o	←	→	o	←

Table 1. In the last stage of the experiment the landmark was either shifted to the right (→), to the left (←) or it remained stationary (o).

second and forth of them with feedback. In these feedback trials an additional small window was presented in the upper right corner of the screen where subjects could see a birdseye view of their path. Little balls indicated the starting– and turning points.

In the last stage a red cone was placed in the vicinity of the starting point. Because of this location this cone was visible to the subjects only at the beginning of a trial and then again during their return to the starting point. When the subjects finished their homevector they passed the landmark already, thus it was not longer visible to them. Each stage consisted of 18 trials (3 length × 6 turning angles) in a random order. During the landmark stage in twelve of these trials the landmark was shifted either to the right or the or the left after the subject had passed the landmark (see table 1). In the remaining six trials the landmark remained at their original position. The whole experiment lasted around an hour.

3 Results

Subjects were able to discriminate between different triangles in this study – i.e. there is a positive correlation between a correct length of a certain homevector and the length that was estimated by the subjects. The same is true for the angle (see figure 3). The subjects could profit from the training in that way that they improved their turning precision.

If a landmark is introduced, the variance of the homing is reduced by a factor[2] of f = 3.4 (figure 4) compared to the results obtained just on the basis of optic flow. This variance reduction is much more pronounced in male (f = 37.89) than in female (f = 2.0) subjects. The latter show also an overall bigger variance in their results than male subjects.

In figure 5 the data of the landmark experiment are plotted. If the landmarks have been shifted, also the homevectors ended in the corresponding direction. This difference is significant, indicating that the landmarks have had a strong influence on this result. The homevectors in the trials with a shifted landmark show less correlation to the "correct" homevector defined by optic flow than to the "correct" homevector defined by the landmarks. They are longer and show a clear overestimation for the turning if the landmark was shifted to the left, respectivly an underestimation for the turning if the landmark was shifted to

[2] fraction of the area of the variance ellipses $f = VAR_{OpticalFlow}/VAR_{landmark}$.

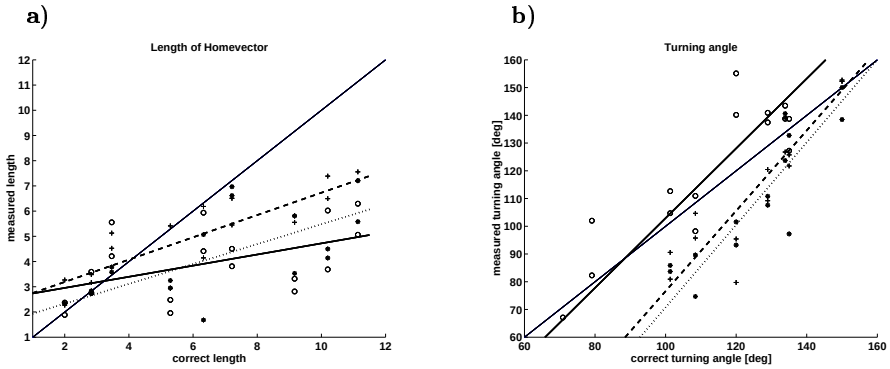


Fig. 3. Linear fit of the length (**a**) and turning angle (**b**) of the homevectors to the correct length and angle; experiments where only optic flow was available (no feedback, no landmark, pooled over 20 subjects). The line style mark the different stages of the experiment: — **o** —: first stage, **- - - * - - -**: second stage, $\cdots + \cdots$: third stage. The THIN line is hypothetical and plotted just for the comparison with error free performance ($f(x) = 1.0x$).

the right (table 2). An indication that also optic flow influenced the homing is the shape of the variance ellipses. If the responses would rely purely on optic flow the variance ellipses for all three cases should be identical, but at locations corresponding to the shifted landmarks. We found that in the trials with shifted landmarks, these ellipses are tilted towards the starting point. This is much more pronounced for the male subjects.

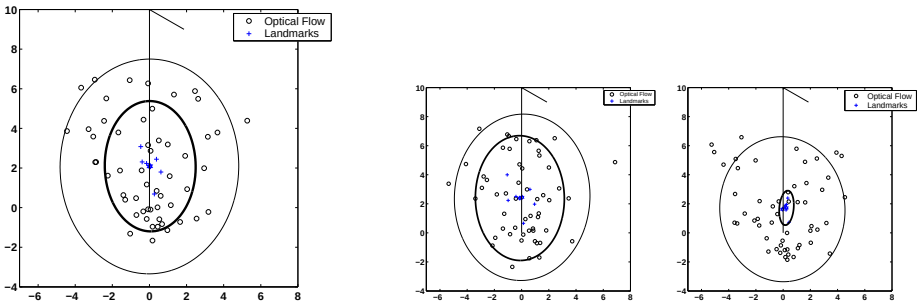


Fig. 4. Endpoints of all hometrajectories and the corresponding variance ellipses. One path of a triangle is added as an example. THIN LINE: Experiments where just optic flow was available. THICK LINE: Homing endpoints if the red cone was placed along the route; data from the trials where the landmark was not pushed. Large Figure: Data of 20 subjects. Middle: Data of 11 female subjects. Right: Data of 9 male subjects.

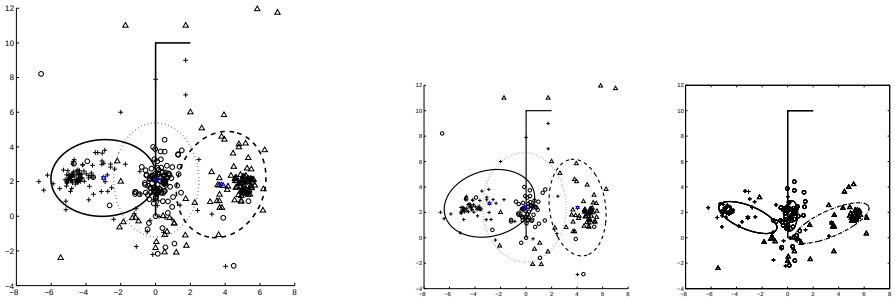


Fig. 5. Endpoint of the hometrajectories (— +: Landmark shiftet to the left; - - - △: Landmark was shifted to the right; · · · **o** : Position of Landmark was constant). Large Figure: Data of 20 subjects. Middle: Data of 11 female subjects. Right: Data of 9 male subjects.

4 Discussion

We studied the integration of optic flow and landmark information. Optic flow in a desktop virtual reality setup is a rather weak cue for path integration, but we were able to train the subjects to make use of optic flow and improve their performance.

The most interesting finding in this experiment is that 19 out of 20 subjects didn't report the transposition of the landmark. On the other hand both types of information were relevant for the behavior of the subjects. This supports the idea that human spatial memory contains isolated chunks of information, which was shown also in experiments performed by Steck et al. [6]. In their experiments subjects could use local and global landmarks for a navigation task. After a training phase the landmarks were transposed. Most of the subjects also didn't report the transpostion, despite the fact that they did use them further for the navigation. In an additional experiment it was shown that their behavior was guided by just one type of landmark, independent of their former prefential landmark type. This shows that the information from the different landmark types was represented in memory but not combined into a single cognitve map.

	stationary landmark	shifted RIGHT	shifted LEFT
optic flow – length difference	1.07 m	0.35 m	-0.747 m
optic flow – turning difference	-7.37^o	44.95^o	-22.37^o
landmark – length difference	1.07 m	1.07 m	1.01 m
landmark – turning difference	-7.37^o	-14.47 ^{o}	5.26^o

Table 2. 1st & 2nd rows: Differences of the measured vectors to the vectors defined by optic flow; **3rd & 4th rows**: Differences of the measured vectors to the homevectors defined by the landmark

In our experiment we find a comparable result for a working memory task. Subjects were able to base their behavior on both types of information but they didn't integrate it in memory – otherwise the tranposition of the landmark should be reported by the subjects. The landmark information dominates the optic flow information. This could be seen by the differences to the corresponding home vectors (Table 2). One reason for this could be, that landmark information resulted in a smaller variance for the homing than optic flow. Similar results have been found in the field of sensor fusion [2]. In a grasping task the information with less variance – which was optic information in this paper in contrast to haptic information – influenced the behavior most.

The difference in our experiment between the male and female subjects is in accordance with the literature, where it is widely accepted that females rely more on landmark navigation than male subjects [5]. For the latter the optic flow should play a more important role. If this would be true in our investigation than the f–values (fraction of variance of homing with optic flow / on the basis of landmarks) should be larger for female than for male subjects. That is definitely not the case. It rather appears that the overall performance of the female subjects is worser compared to the male subjects in our experiments. This might be explained by different amounts of experience with computers and in particular with video games.

It is clear, that the type of the landmark and the position of the landmark have a strong influence on the result. In our experiments we used a rather simple landmark, placed along the route. In further experiments it will be worthwhile to investigate the influence of other types of landmarks or landmark arrays.

References

1. Bertin, RJV, Israël, I, Lappe, M: Perception of two–dimensional, simulated ego–motion trajectories from optic flow. *Vision Reseach* **40**, pp. 2951–2971, 2000
2. Ernst MO, Banks MS: Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, pp. 429–433, 2002
3. Loomis JM, Klatzky RL, Golledge RG, Cicinelli JG, Pellegrino JW, Fry PA.: Nonvisual navigation by blind and sighted: assessmet of path integration ability. Journal of Experimental Psychology: General. **122**, pp. 73–91, 1993
4. Riecke BE, vanVeen HAHC, Bülthoff HH: Visual Homing is possible without Landmarks – A Path Integration Study in Virtual Reality. Perception, in press, 2002
5. Sandstrom NJ, Kaufman J, Huettel SA: Males and females use different distal cues in a virtual environment navigation task. Cognitive Brain Research,**6**, pp 351–360, 1998
6. Steck S, Mallot HA: The role of global and local landmarks in virtual environment navigation. Presence. Teleoperators and Virtual Environments. **9**, pp. 69 –83, 2000
7. Warren WHJ, Blackwell AW, Kurtz KJ, Hatsopoulos NG & Kalish ML: On the sufficiency of the velocitiy field for perception of heading, Biological Cybernetics, **65**, pp. 311 – 320, 1991

Simultaneous Estimation of Extended Optical Flow and Global Parameters

Moritz Diehl [1], Ralf Küsters [1,2] and Hanno Scharr [1,2]

[1] Interdisciplinary Center for Scientific Computing, Ruprecht Karls University, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany
[2] Institute for Chemistry and Dynamics of the Geosphere, Institute III: Phytosphere, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany
{Moritz.Diehl,Ralf.Kuesters,Hanno.Scharr}@iwr.uni-heidelberg.de

Abstract. A novel algorithm for simultaneous estimation of many local and a few global parameters in image sequences is presented. Usual parameter estimation frameworks as e.g. the structure tensor method for extended optical flow [8] are designed for local parameters only. There the estimation can be performed for every pixel neighborhood separately. Global parameters effect a full coupling of the model equation matrix. The main idea in this paper is to split the model equation matrix into an easily invertible local parameter part and a small global parameter part. We compare our new approach to two common estimation methods. In a performance evaluation of systematic errors and noise stability the superior behaviour of the new approach is demonstrated.

Keywords: extended optical flow, least squares parameter estimation, large-scale optimization, image sequences

1 Introduction

Combined motion and brightness change estimation in physically motivated models proved to be successful in many applications (e.g.[15, 16, 10]). In well established parameter estimation frameworks as e.g. the structure tensor method (total least squares (TLS) approach) [9, 8] or its mixed ordinary least squares (OLS) and TLS version [6] physical models with *local* parameters only can be applied. These methods can be implemented efficiently in terms of RAM needed and CPU time used as all estimations can be performed separately for each pixel neighborhood. In other words, the model equation matrix is a block diagonal matrix with one block per pixel and we process one block after the other. This is no longer true if global parameters have to be estimated as well. They introduce full rows in the model matrix, thus coupling all blocks. In this paper we present an OLS estimation method for simultaneous estimation of local and global parameters. It has comparable complexity and memory requirements as pure local methods.

The example model used here is designed for optical flow estimation where the camera has an automatic gain control. This is the case for most consumer camcorders and thus a quite interesting application.

Related work. Although there is a rich literature on optical flow estimation techniques (see [12, 1] for current overviews), direct extensions have been studied

to a much smaller extent. There are extensions towards affine motion estimation [3, 4], flow in texture and depth maps [18], physically motivated brightness changes [11] and robust estimations [6, 2]. Regularization schemes [19], special filters [17, 14, 5] and coupled denoising methods [20] have been developed. But to the best of our knowledge there is no extension using global parameters as e.g. camera gain.

2 The Model

An automatic gain control changes gray values $g(x, y, t)$ in space-time by

$$\frac{\mathrm{d}g(x,y,t)}{\mathrm{d}t} = k(t)g(x,y,t)$$

where $k(t)$ a spatially constant factor describing gain changes. For optical flow estimation we get for each pixel one model equation

$$\begin{aligned} \tfrac{\mathrm{d}g(x,y,t)}{\mathrm{d}t} = \tfrac{\partial g}{\partial x}\tfrac{\mathrm{d}x}{\mathrm{d}t} + \tfrac{\partial g}{\partial y}\tfrac{\mathrm{d}y}{\mathrm{d}t} + \tfrac{\partial g}{\partial t} &= k(t)g(x,y,t) \\ \Leftrightarrow \quad g_x u_x + g_y u_y + g_t &= kg \end{aligned}$$

using the notation $g_* = \frac{\partial g}{\partial *}$ and substituting $u_* = \frac{\mathrm{d}*}{\mathrm{d}t}$. The local parameters are the motion components u_x and u_y, the global parameter is the gain factor k. Let us in the following only consider the central image of a temporal slice of the sequence, and order the $N = N_y \times N_x$ pixels of the image in some arbitrary way, numbered with an index $i = 1, \ldots, N$, replacing the space coordinates. Given the coefficients g_x^i, g_y^i, g_t^i, and g^i at each pixel (e.g. using the derivative convolution kernels given in [17, 14, 5]), we want to determine an estimate for the local parameters $u_x^1, u_y^1, u_x^2, u_y^2, \ldots, u_x^N, u_y^N$ and for the global parameter k, that is best in a least squares sense. For this aim, we define a neighborhood Ω_i (with n_Ω pixels) around each pixel i. Then we define a least squares term $\sum_{j\in\Omega_i}(g_x^j u_x^i + g_y^j u_y^i + g_t^j - kg^j)^2$, which measures the misfit of estimated parameters and image data in each neighborhood. The approach followed in this paper is to minimize the sum of all misfits:

$$\sum_{i=1}^{N} \sum_{j\in\Omega_i} (g_x^j u_x^i + g_y^j u_y^i + g_t^j - kg^j)^2 \tag{1}$$

by varying the parameters u_x^i, u_y^i, $i = 1, \ldots, N$, and k. If the global gain parameter k was *not* present ($k = 0$), the optimization could be carried out pixelwise for $i = 1, \ldots, N$, thus allowing for an efficient sequential processing over the whole data set. The same applies for a local gain estimation, where k is replaced by a *local* gain factor k^i in each local misfit term. In our case, with a *global* gain parameter k, a coupling between all terms is introduced. Thus, the above optimization problem has to be treated as a large scale problem and can only be solved for practical problems, if the problem structure is carefully exploited. In this paper, we propose a numerical solution method which achieves this aim by

making use of the so called *Sherman-Morrison-Woodbury-Formula*, which allows to efficiently obtain the inverse of an easily invertible matrix when it is modified by a low rank matrix.

3 The Novel Algorithm

The idea is as follows: defining the parameter vector $x = (u_x^1, u_y^1, u_x^2, u_y^2, \ldots, u_x^N, u_y^N, k)^T$, $x \in \mathbb{R}^n$, $n = 2N + 1$, the large scale optimization problem with objective (1) can be summarized in the form $\min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2$ and the solution vector $\bar{x}$ necessarily satisfies the *normal equation*

$$A^T A\bar{x} = A^T b, \tag{2}$$

i.e., $\bar{x} = (A^T A)^{-1} A^T b$, if $A^T A$ is invertible. The matrix A has the following block structure

$$A = \left[\begin{array}{cccc|c} B_1 & & & & V_1 \\ & B_2 & & & V_2 \\ & & \ddots & & \vdots \\ & & & B_N & V_N \end{array}\right] = \left[\, B \middle| V \,\right] \tag{3}$$

with $n_\Omega \times N_{\text{lp}}$-blocks B_i and $n_\Omega \times N_{\text{gp}}$-blocks V_i. In the above application, we have $N_{\text{lp}} = 2$ and $N_{\text{gp}} = 1$. The squared system matrix consequently has the form

$$A^T A = \left[\begin{array}{c|c} B^T B & B^T V \\ \hline V^T B & V^T V \end{array}\right].$$

Each of these matrix products can be calculated efficiently using convolutions (compare [13, 15] for the TLS case). Finally, the squared matrix can be decomposed as

$$A^T A = M + RSR^T$$

with

$$M = \left[\begin{array}{c|c} B^T B & 0 \\ \hline 0 & V^T V \end{array}\right],\ R = \left[\begin{array}{c|c} B^T V & 0 \\ \hline 0 & \mathbb{I} \end{array}\right],\ S = \left[\begin{array}{c|c} 0 & \mathbb{I} \\ \hline \mathbb{I} & 0 \end{array}\right],$$

where M is block diagonal and R is a matrix of low rank, $2N_{\text{gp}}$, so that the Sherman-Morrison-Woodbury formula [7] can be used to efficiently compute the inverse:

$$(A^T A)^{-1} = M^{-1} - M^{-1}R(S^{-1} + R^T M^{-1} R)^{-1} R^T M^{-1}.$$

In addition to the matrix blocks $B_i^T B_i$ and $\sum_{i=1}^N V_i^T V_i$ of M we therefore only have to invert one further $(2N_{\text{gp}}) \times (2N_{\text{gp}})$ matrix, $(S^{-1} + R^T M^{-1} R)$, and all remaining calculations for computation of

$$\bar{x} = (A^T A)^{-1} A^T b = \left(\mathbb{I} - M^{-1}R(S^{-1} + R^T M^{-1} R)^{-1} R^T\right) M^{-1} A^T b$$

can be performed as matrix vector products. As the inversion of the matrix blocks $B_i^T B_i$ is by far the most time consuming step in the computations of the

algorithm, the computational burden is comparable to that of an OLS velocity estimation *without* gain estimation. When *local* gains k^i are estimated, this results in a completely decoupled problem, but with larger local matrix blocks, so that the computational burden is considerably higher than for the proposed approach.

4 Experimental Validation and Comparison with Existing Methods

In order to quantify the accuracy and noise stability we measure the velocity of a translating "wave"-pattern with a global brightness change $g(x, y, t) = \exp(kt)\cos(2\pi(x - u_x t)/\lambda_x) * \cos(2\pi(y - u_y t)/\lambda_y)$ with varying wave lengths λ_x, λ_y. The camera gain factor is here given by $\exp(kt)$, and k is constant in time. To those sequences, normal distributed noise with a standard deviation up to $\sigma = 10\%$ is added. For comparison, we calculated the velocities u_x and u_y using three estimation models: the first does not estimate the gain ("no gain", $k = 0$), the second estimates local gain factors k^i, $i = 1, \ldots, N$ ("local gain"), and the third is our new algorithm estimating a spatially global gain k for each picture ("global gain"). The first two algorithms use the well known OLS method for local parameters (see e.g. [12]). Below some results of these tests are shown.

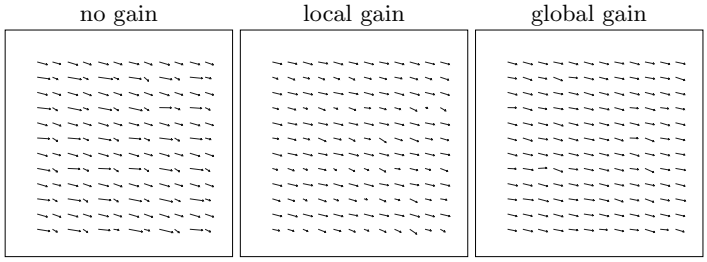


Fig. 1. Estimated flowfields for three different models.

The picture shows the estimated velocity vectors for the three estimation methods, using a a simulation with 5 % gain and 5% noise. It can be seen that the velocity estimates of the first model, which is not able to capture the gain, are highly distorted. The velocity estimates of the global gain model show less variation than the local gain model. This is due to the fact that the effects of noise are better dampened out by inclusion of the knowledge that k is spatially constant. The corresponding variances in u_x and u_z are shown in the second last line of the table below, which also shows variances for some other gain and noise scenarios.

As expected, all models capture well the scenario without gain and without noise, whereas the no gain model has increasing difficulty with growing gains in the data. Compared to the local gain model, the global gain model shows

Fig. 2. Variances and relative estimation errors of the velocity estimates, tested for different scenarios with varying gain k and noise level σ, using different models

		Variances $[10^{-4}]$						Rel. est. errors $[10^{-3}]$					
k [%]	σ [%]	no gain		local gain		global gain		no gain		local gain		global gain	
		u_x	u_y	u_x	u_y	u_x	u_y	u_x	u_y	u_x	u_y	u_x	u_y
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	5	56	71	143	125	58	72	56	121	103	232	57	122
0	10	283	349	531	266	251	308	172	676	253	607	163	590
1	0	19	31	0	0	0	0	0	50	0	0	0	0
1	5	75	99	141	117	64	77	58	169	101	211	59	131
1	10	263	335	514	280	256	311	165	641	254	625	165	603
5	0	477	763	0	0	0	0	0	1205	0	0	0	0
5	5	483	742	148	119	56	71	57	1262	102	217	57	121
5	10	587	816	523	270	262	324	157	1553	249	604	167	634

comparable or lower variances. Similar observations hold for the following table, where the mean relative errors in the velocity estimates (compared to the correct values) are listed for the same set of scenarios.

Note that the computational load of the proposed algorithm for global gain estimation is *smaller* than that of the OLS method for local gain estimation, because far less free parameters have to be determined.

5 Summary and Outlook

We have introduced a novel algorithm for simultaneous estimation of many local and a few global parameters in image sequences. A numerically efficient algorithm to solve the arising large scale least squares optimization problems is presented. The algorithm is based on the idea to split the model equation matrix into an easily invertible local parameter part and a low rank part introduced by the presence of global parameters. The inversion of the combined system is efficiently performed by means of the so called Sherman-Morrison-Woodbury-Formula. The resulting algorithm has comparable complexity and memory requirements as a pure local method without estimation of the global parameters.

The capacity of the new algorithm to cope with global gains is demonstrated in a first series of numerical experiments. The resulting velocity estimates compare well with those obtained by existing OLS methods with local gain estimation, and are affected less by noise.

Further work will focus on extending the simultaneous local-global parameter estimation towards TLS formulations, and on replacing the normal equation approach (2) by a suitable, structure exploiting Q-R factorization of the system matrix (3).

References

1. J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. In *IJCV*, pages 43–77, 1994. 12(1).
2. M. Black, D. Fleet, and Y. Yacoob. Robustly estimating changes in image appearence. *CVIU*, 7(1):8–31, 2000.
3. G. Farnebäck. Fast and accurate motion estimation using orientation tensors and parametric motion models. In *ICPR*, pages 135–139, 2000.
4. D. Fleet. *Measurement of Image Velocity.* Kluwer, 1992.
5. D.J. Fleet and K. Langley. Recursive filters for optical flow. *IEEE Trans. PAMI*, 17(1):61–67, January 1995.
6. C. Garbe. *Measuring Heat Exchange Processes at the Air-Water Interface from Thermographic Image Sequence Analysis.* PhD thesis, Heidelberg University, 2001.
7. G. H. Golub and C. F. van Loan. *Matrix Computations.* Johns Hopkins University Press, 3 edition, 1996.
8. H. Hauecker and D.J. Fleet. Computing optical flow with physical models of brightness variation. *PAMI*, 23(6):661–673, June 2001.
9. H. Hauecker, C. Garbe, H. Spies, and B. Jhne. A total least squares for low-level analysis of dynamic scenes and processes. In *DAGM 1999*, 1999. 240–249.
10. H. Hauecker, U. Schimpf, C.S. Garbe, and B. Jhne. Physics from IR image sequences: Quantitative analysis of transport models and parameters of air-sea gas transfer. In *Gas Transfer at Water Surfaces.* American Geophysical Union, 2001.
11. H. Haußecker and D. Fleet. Computing optical flow with physical models of brightness variation. In *CVPR*, 2000.
12. H. Haußecker and H. Spies. Motion. In *Handbook of Computer Vision and Applications.* Academic Press, 1999.
13. B. Jähne. Performance characteristics of low-level motion estimators in spatiotemporal images. In W. Foerstner, editor, *DAGM-Workshop Performance Characteristics and Quality of Computer Vision Algorithms*, Univ. Bonn, 1997.
14. B. Jähne, H. Scharr, and S. Körkel. Principles of filter design. In *Handbook of Computer Vision and Applications.* Academic Press, 1999.
15. B. Jhne, H. Hauecker, H. Scharr, H. Spies, D. Schmundt, and U. Schurr. Study of dynamical processes with tensor-based spatiotemporal image processing techniques. In *ECCV 1998*, pages 322–336. Springer, 1998.
16. N. Kirchgener, H. Spies, H. Scharr, and U. Schurr. Root growth analysis in physiological coordinates. In *ICIAP'01*, Palermo, Italy, 2001.
17. H. Scharr. *Optimal Operators in Digital Image Processing.* PhD thesis, University of Heidelberg, Germany, 2000.
18. H. Spies, H. Haußecker, B. Jähne, and J.L. Barron. Differential range flow estimation. In *DAGM*, pages 309–316, 1999.
19. H. Spies, N. Kirchgener, H. Scharr, and B. Jhne. Dense structure estimation via regularised optical flow. In *VMV 2000*, pages 57–64, Saarbrücken, Germany, 2000.
20. H. Spies and H. Scharr. Accurate optical flow in noisy image sequences. In *ICCV'01*, pages 587–592, Vancuver, Canada, 2001.

Dynamical retino-cortical mapping

Markus A. Dahlem and Florentin Wörgötter

Computational Neuroscience
Department of Psychology
University of Stirling
Stirling FK9 4LA
Scotland / UK
{mad1, faw1}@cn.stir.ac.uk
http://www.cn.stir.ac.uk/

Abstract. A dynamical mapping strategy is introduced, that leads to a novel representation of optical flow in which motion parallax depth cues are reliably obtained. It is known that similar data preprocessing is performed for visuomotor control tasks, and two dynamic mapping versions are advocated by various groups; either mapping into eye-centered coordinates, or into head-, and body coordinates. While for remembered target locations each of these mappings has its specific advantages, we show here, that optical flow is only simplified when dynamically mapped into head coordinates. There is little if any benefit for a depth-from-motion algorithm in a dynamic retinotopic map. Our results can be utilized in technical visual systems and we also suggest a verifiable hypothesis about a such a representation of optical flow in extrastriate cortex.

1 Introduction

One of the chief problems in computational vision is the three-dimensional reconstruction of a static scene from two-dimensional images [1]. Motion parallax is one of the depth cues that can be used to recover the three-dimensional structure of a viewed scene [2]. Motion induces a velocity field on the retina called the optical flow [3]. In the most general motion case, i. e., ego- plus object motion, the resulting curved optical flow field pattern cannot be resolved for depth analysis without additional assumptions [4] and even if simplifying assumptions are made, the problem of depth-from-motion remains rather complex.

Purely translational ego-motion induces one of the simplest optical flow fields. The optical flow has a fixed point, called the focus of expansion (FOE). All optical flow trajectories move outwards from the FOE. A radial flow field (RFF) contains reliable and rather easily accessible information about the three-dimensional structure of the viewed scene [5]. It is readily seen that the motion in such an RFF is one-dimensional in any retinotopic map—in a specific curve-linear coordinate system. For example, in retinal coordinates the RFF is expanding solely along the radial coordinate when an observer approaches an object. In coordinates of primate striate cortex this radial flow is mapped roughly along

parallel aligned neurons starting from the posterior pole to more anterior location in the medial occipital lobe [6]. We show in this study, that a flow field, with the only rotational components due to eye-gaze movements, is in a dynamic map isometric to a one-dimensional RFF. In other words, this flow filed is invariant to these specific rotational components. There are areas in extrastriate cortex known to have similar features as the dynamic map we suggest.

One of us (FW) introduced earlier an algorithm that efficiently analyzes an RFF, i.e., purely translational flow, and then reconstructs the viewed three-dimensional scene [5]. Details of the algorithm should be taken from that reference. We will use this algorithm to explore the use of a dynamic map—as described in the next section— for ego-motions with eye-gaze movements combined with straight body motion. We would like to emphasize, that any other depth-from-motion algorithms, that takes as input an one-dimensional RFF, can utilize the dynamical mapping strategy. However, the RFF-algorithm has been specifically designed to allow for parallelization of computations, and it is foremost this feature which is conserved by dynamic mapping.

2 Dynamical Mapping

To map the retinal flow field to a head centric frame, the retina is sampled by point-like receptive fields (Fig. 1, top layer). Initially, the receptive fields are placed such that they sample an RFF where direction of gaze and heading direction coincide. The receptive fields are positioned on a polar grid (receptive field grid, RFG) defined by radial axes expanding from the FOE. If the distance between successive receptive fields increases hyperbolically on each radial line, the optical flow is sampled uniformly.

The layout of the receptive fields on the RFG matches the radial optical flow field only if motion direction and direction of gaze coincide. When both directions differ by a constant angle α the receptive field positions on the RFG are re-mapped. After a gaze shift α about the Y-axis (angle of yaw), the optical flow is transformed by:

$$\theta^{hc}(\alpha) = f\sqrt{\left(\frac{\theta\cos(\phi)\cos\alpha - f\sin\alpha}{f\cos\alpha + \theta\cos(\phi)\sin\alpha}\right)^2 + \left(\frac{\theta\sin(\phi)}{f\cos\alpha + \theta\cos(\phi)\sin\alpha}\right)^2} \quad (1)$$

$$\phi^{hc}(\alpha) = \arctan\left(\frac{\theta\sin(\phi)}{\theta\cos(\phi)\cos\alpha - f\sin\alpha}\right)$$

The index hc indicates that these coordinates are head-centric while without index they are retinotopic.

In a head-centric frame the rotational component of the optical flow is counteracted by constantly updating the receptive field positions along with the rotational component according to the mapping function (Eq. 1). This is possible because direction and magnitude of the rotational component depend only on the angular velocity of the gaze change and not on any external information of the viewed scene. The optical flow in a head-centric frame is then congruent to

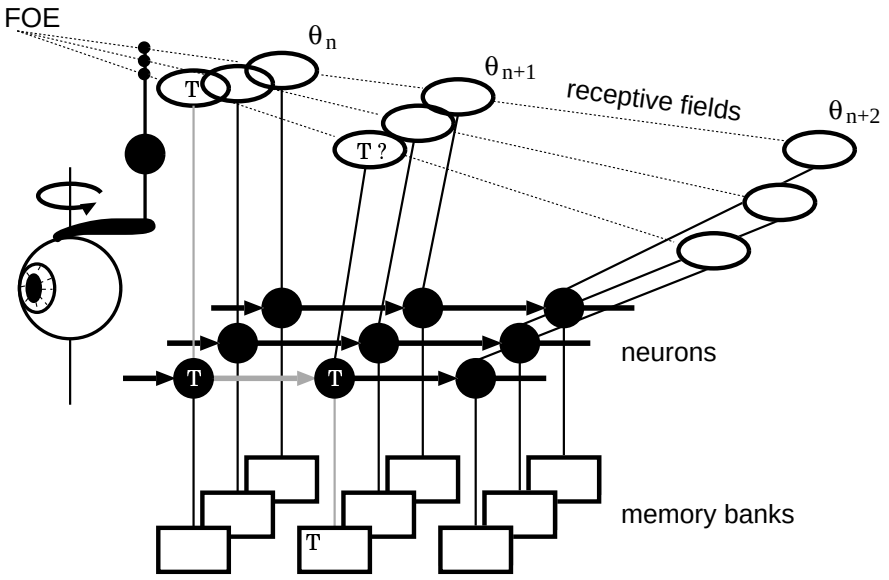


Fig. 1. Architecture of the three layer network. The top layer consists of receptive fields sampling the optical flow. Each receptive field projects to a neuron in the middle layer. The third layer consists of memory banks, one for each processing neuron. A separate neuron represents a structure mapping eye-positions. A visual tokens (T) is passed from the receptive field along the exemplarily shown grey connections towards the memory bank of a consecutive neuron. A head-centric representation of visual input in the middle neuronal layer is achieved by dynamically mapping the receptive field positions according to the direction of gaze. To re-construct three-dimensional position of viewed objects, the middle layer needs only locally exchanged information in one spatial direction (from left to right).

an RFF obtained with stable direction of gaze. Consequently, this dynamic map is invariant under eye-gaze movements and the RFF-algorithm can be applied on this head-centric map.

3 Performance of the RFF-algorithm on a head-centric map

An observer moving straight without changing the direction of gaze can adequately detect the three-dimensional position of the edges of objects in view by the RFF-algorithm [5]. For example, determining the distance of a teapot by the RFF-algorithm, results in three-dimensional coordinates, shown in front view (Fig. 2 A) and top view (Fig. 2 B). These detected coordinates outline the contour of the teapot. The depth coordinate Z, as shown in the top view (Fig. 2 B), is the actual output of the RFF-algorithm. The contour in the other

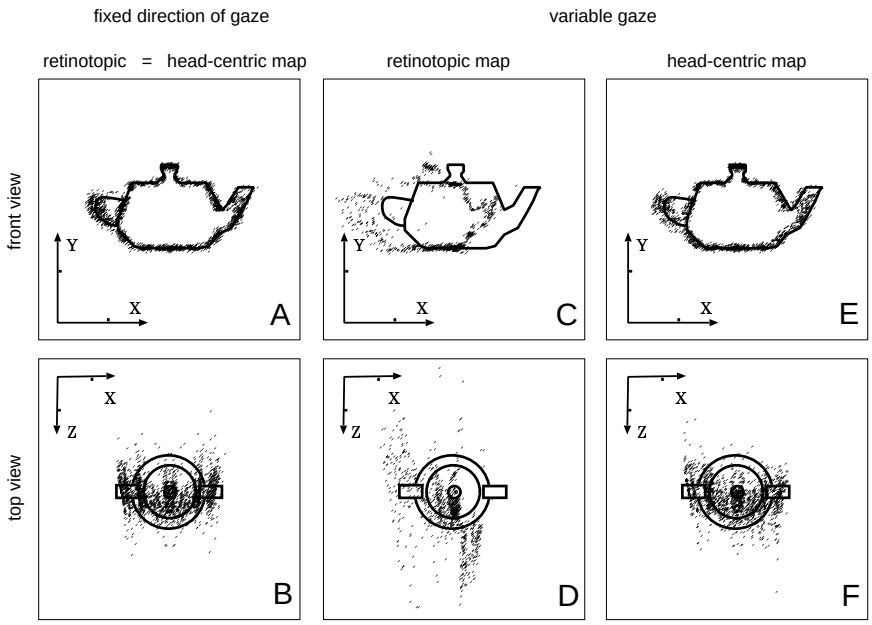


Fig. 2. A teapot viewed with stable and variable gaze. The position of the teapot in Cartesian coordinates (X,Y,Z) can be detected on a retinotopic map by the RFF-algorithm only when the gaze is pointing toward a fixed direction (A and B). Otherwise this algorithm makes systematic errors (C and D). If the position of the teapot is to remain stable, this algorithm must operate on a head-centric map (E and F). See also text.

two coordinates, X and Y (Fig. 2 A), are directly projected onto the retina and therefore they are already implicitly known, except for a scaling constant.

The detection of the teapot deteriorates when the straight body motion is combined with eye-gaze movements (front view Fig. 2 C, and top view D). There is even a shift of the projection of the teapot in the X-direction, that is, in the direction of one implicitly known coordinate (Fig. 2 C). This shift is inherent in the retino-centric map. Such a map can not statically store spatial locations. To be precise, edges of the teapot that are located on the retina right (left) from the FOE are accelerated (slowed down) by the additional rotational flow component, when the gaze rotates clock-wise about the Y-axis. This systematic change in the flow velocity is falsely interpreted by the RFF-algorithm as an edge too near (far), as shown by the tilt in Fig. 2 D. If the RFF-algorithm operates on head-centric optical flow fields, the performance of the RFF-algorithm is invariant under gaze sifts. (Fig. 2 E and F).

To quantify the performance of the RFF-algorithm on both the retinal flow field and the head-centric flow field, we defined a standard detection task: the three-dimensional reconstruction of a centric viewed square plane. For fixed direction of gaze along heading direction this corresponds to a situation where

edges move with hyperbolically increasing velocity along the receptive fields on each radial line. The angles between the edge and the radius vary between 0° and 45°. The average error in the detected three-dimensional position of the edges of the square plane was normalized to 1 for fixed direction of gaze (Fig. 3). If the gaze direction rotates stepwise by a total angle between 1° and 4° about the Y-axis, the error increases when the RFF-algorithm operates on retinal optical flow fields, as expected (see Fig 3). On head-centric optical flow fields the performance of the standard detection task is stable.

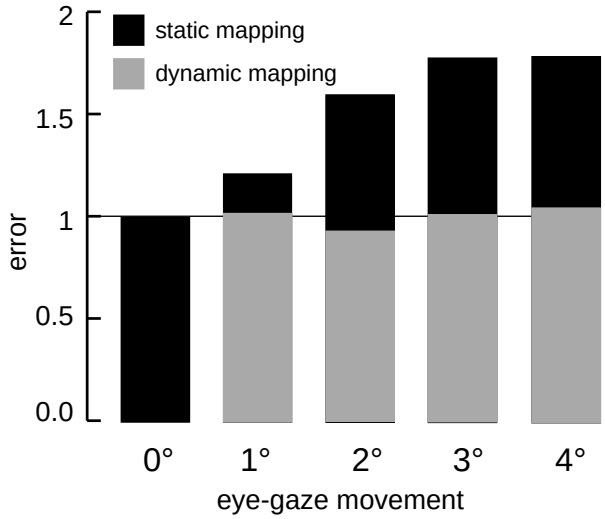


Fig. 3. Performance of the RFF-algorithm operating on a retinotopic map compared to a head-centric map. While on a head-centric map the performance is stable, on the retinotopic map it fastly deteriorates.

4 Discussion

Rotational components, foremost in form of smooth pursuit eye movements, are likely to occur in the ego-motion even within short periods of time. As soon as a rotational component is mixed with translation motion, the optical flow is two-dimensional in any coordinate system of a retinotopic map and extracting depth from optical flow becomes generally far more complicated. We showed that with a simple dynamic mapping strategy of visual space, the effect of eye-gaze movements on the optical flow can be eliminated. The resulting flow field on a head-centric map is congruent to the one induced by pure translational motion.

Dynamical mapping provides an example of combining two visual brain maps into one. In this case a subcortical sensor map that controls gaze direction in

retinal coordinates [7] and a retinotopic cortical map. The resulting map has qualitative new and advantageous features. We also attach importance to a head-centric map because it serves multiple though related purposes. In several areas in the parietal cortex: V3a [8], V5 [9], MST [10], V6 [11], V6a [12], 7a [13], and VIP [14] the activity of neurons is influenced by gaze direction. Precise gaze tuning together with a topographic representation of space can form a head-centric map. Area MST [15] and 7a [16] are both known to represent optical flow, although in different ways, and are likely candidates to utilize one-dimensional flow fields as depth cues, as we suggest here. To test this hypothesis, one needs to present radial expanding optical flow and introduce rotational components by pursuit gaze movements.

References

1. Marr, D. (1982). Vision. New York: W. H. Freeman and Company.
2. Nakayama, K., & Loomis, J. M. (1974). Optical velocity patterns, velocity-sensitive neurons, and space perception: a hypothesis. *Perception*, 3, 63–80.
3. Gibson, J. J. (1950). The perception of the visual world. Boston: Houghton Mifflin.
4. Poggio G.F., Torre V., Koch C.: Computational vision and regularization theory. Nature, **317** (1985) 314–319.
5. Wörgötter, F., Cozzi, A., & Gerdes V. (1999). A parallel noise-robust algorithm to recover depth information from radial flow fields. *Neural Computation*, 11, 381–416.
6. Tootell, R. B., Silverman, M. S., Switkes, E., & De Valois, R. L. (1982). Deoxyglucose analysis of retinotopic organization in primate striate cortex. *Science*, 218, 902–904.
7. Klier, E.M., Wang, H., & Crawford, J. D. (2001). The superior colliculus encodes gaze commands in retinal coordinates. *Nature Neuroscience*, 6, 4 627-632.
8. Galletti, C., & Battaglini P. P. (1989). Gaze-dependent visual neurons in area V3A of monkey prestriate cortex. *Journal of Neuroscience*, 9, 1112–1125.
9. Puce, A., Allison, T., Bentin, S., Gore, & J. C., McCarthy G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience*, 18, 2188–2199.
10. Squatrito, S, & Maioli, M. G. (1996). Gaze field properties of eye position neurones in areas MST and 7a of the macaque monkey. *Visual Neuroscience*, 13, 385–398.
11. Galletti, C., Battaglini, P. P., & Fattori P. (1995). Eye position influence on the parieto-occipital area PO (V6) of the macaque monkey. *European Jornal of Neuroscience*, 7, 2486–2501.
12. Galletti, C., Fattori, P., Battaglini, P. P., Shipp, S, & Zeki, S. (1996). Functional demarcation of a border between areas V6 and V6A in the superior parietal gyrus of the macaque monkey. *European Jornal of Neuroscience* 8, 30–52.
13. Andersen, R. A., Essick G. K., & Siegel, R. M. (1985). Encoding of spatial location by posterior parietal neurons. *Science*, 230, 456–458.
14. Duhamel, J. R., Bremmer, F., BenHamed, S., & Graf, W. (1997). Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature*, 389, 845–848.
15. Duffy, C. J., & Wurtz, R. H. (1995). Response of monkey M. ST neurons to optic flow stimuli with shifted centers of motion. *Journal of Neuroscience*, 15, 5192–5208.
16. Siegel, R. M., Read, H. L. (1997). Analysis of optic flow in the monkey parietal area 7a. *Cerebral Cortex* 7, 327–346.

A Neurally-Inspired Model for Detecting and Localizing Simple Motion Patterns in Image Sequences

Marc Pomplun[1], Yueju Liu[2], Julio Martinez-Trujillo[2], Evgueni Simine[2], and John K. Tsotsos[2]

[1]Department of Computer Science, University of Massachusetts at Boston, Boston, MA 02125, USA
[2]Centre for Vision Research, York University, Toronto, Canada M3J 1P3

Abstract. In the present paper, we propose a neurally-inspired model of the primate motion processing hierarchy and describe its implementation as a computer simulation. The model aims to explain how a hierarchical feedforward network consisting of neurons in the cortical areas V1, MT, MST, and 7a of primates achieves the detection of different kinds of motion patterns. Moreover, the model includes a feedback gating network that implements a biologically plausible mechanism of visual attention. This mechanism is used for sequential localization and fine-grained inspection of every motion pattern detected in the visual scene.

1 The Feedforward Mechanism of Motion Detection

In the present paper, we propose a neurally-inspired model of the primate motion processing hierarchy and describe its implementation as a computer simulation. The model aims to explain how a hierarchical feed-forward network consisting of neurons in the cortical areas V1, MT, MST, and 7a of primates achieves the detection of different kinds of motion patterns.

Cells in *striate area V1* are well known to be tuned towards a particular local speed and direction of motion in at least three main speed ranges [1]. In the model, V1 neurons estimate local speed and direction in five-frame, 256×256 pixel image sequences using spatiotemporal filters (e.g., [2]). Their direction selectivity is restricted to 12 distinct, Gaussian-shaped tuning curves. Each tuning curve has a standard deviation of 30º and represents the selectivity for one of 12 different directions spaced 30º apart (0º, 30º, …, 330º). V1 is represented by a 60×60 array of hypercolumns. The receptive fields (RFs) of V1 neurons are circular and homogeneously distributed across the visual field, with RFs of neighboring hypercolumns overlapping by 20%.

In *area MT* a high proportion of cells are tuned towards a particular local speed and direction of movement, similar to direction and speed selective cells in V1 [3, 4]. A proportion of MT neurons are also selective for a particular angle between movement direction and spatial speed gradient [5]. Both types of neurons are represented in the MT layer of the model, which is a 30×30 array of hypercolumns. Each MT cell receives input from a 4×4 field of V1 neurons with the same direction and speed selectivity.

Neurons in *area MST* are tuned to complex motion patterns: expand or approach, shrink or recede, rotation, with RFs covering most of the visual field [6, 7]. Two types of neurons are modeled: one type selective for translation (as in V1) and another type selective for spiral motion (clockwise and counterclockwise rotation, expansion, contraction and combinations). MST is simulated as a 5×5 array of hypercolumns. Each MST cell receives input from a large group (covering 60% of the visual field) of MT neurons that respond to a particular motion/gradient angle. Any coherent motion/gradient angle indicates a particular type of spiral motion.

Finally, *area 7a* seems to involve at least four different types of computations [8]. Here, neurons are selective for translation and spiral motion as in MST, but they have even larger RFs. They are also selective for rotation (regardless of direction) and radial motion (regardless of direction). In the simulation, area 7a is represented by a 4×4 array of hypercolumns. Each 7a cell receives input from a 4×4 field of MST neurons that have the relevant tuning. Rotation cells and radial motion cells only receive input from MST neurons that respond to spiral motion involving any rotation or any radial motion, respectively.

Fig. 1 shows the activation of neurons in the model as induced by a sample stimulus. Note that in the actual visualization different colors indicate the response to particular angles between motion and speed gradient in MT gradient neurons. In the present example, the gray levels indicate that the neurons selective for a 90º angle gave by far the strongest responses. A consistent 90º angle across all directions of motion signifies a pattern of clockwise rotation. Correspondingly, the maximum activation of the spiral neurons in areas MST and 7a corresponds to the clockwise rotation pattern (90º angle). Finally, area 7a also shows a substantial response to rotation in the medium-speed range, while there is no visible activation that would indicate radial motion.

2 The Feedback Mechanism of Visual Attention

Most of the computational models of primate motion perception that have been proposed concentrate on bottom-up processing and do not address attentional issues. However, there is evidence that the responses of neurons in areas MT and MST can be modulated by attention (Treue & Maunsell, 1996). Moreover, we claim that attention is necessary for a precise localization of motion patterns in image sequences. As a result of the model's feedforward computations, the neural responses in the high-level areas (MST and 7a) roughly indicate the kind of motion patterns presented as an input but do not localize the spatial position of the patterns.

In order to create a comprehensive motion model that is in agreement with biological findings and is capable of localizing motion patterns, we added a mechanism of visual attention to it. We decided to use the biologically plausible Selective Tuning approach [9], requiring the introduction of a feedback gating network to the model. Each neuron in the original motion hierarchy received an assembly of gating units that control the bottom-up information flow to that neuron.

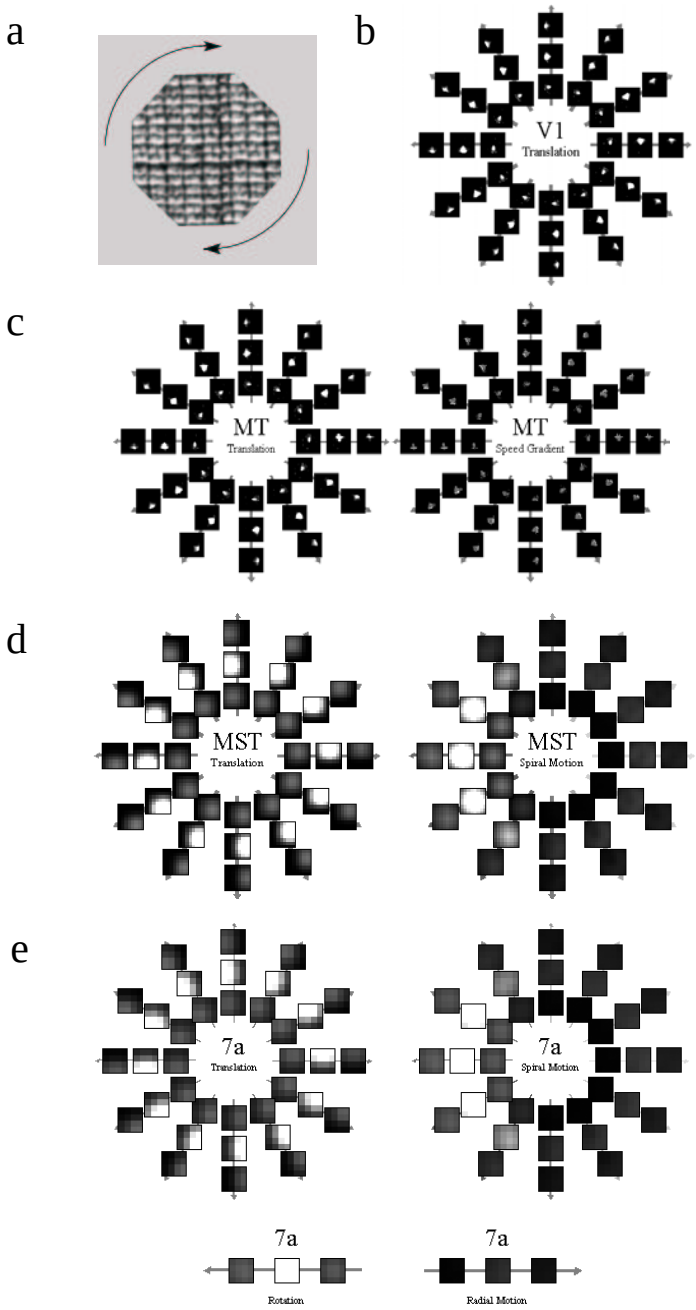


Fig. 1. The model's response to a clockwise rotating stimulus (panel a). Brightness indicates activation in areas V1, MT, MST, and 7a (panels b to e). Arrows represent selectivity for direction of motion or the angle between motion and speed gradient, and the three concentric circles stand for the three speed selectivity ranges in the model.

The attentional processing works as follows: First, a “motion activity" map with the same size as a 7a layer is constructed after the bottom-up processing. The value of a node in the activity map is a weighted sum of the activations of all 7a neurons at this position and it reflects the overall activation. Second, a WTA (Winner-Take-All) algorithm finds the globally most active location. Then at this location, two WTAs will compete among all the translational motion patterns and spiral motion patterns respectively and thus result in two winner neurons. A WTA runs among the winners' gating units, whose activation pattern is initially identical to the one in the winner neurons' RFs. The resulting winners activate the connected neurons in lower layers, whereas the bottom-up information flow through the losing gating units is inhibited. This process continues until the bottom layer, and the recognized motions are localized in the input sequence. The gating network then inhibits the feed-forward processing of neighboring motion patterns so that no interfering information reaches the higher levels of the model. Loosely speaking, the model “focuses its attention“ on the winning motion pattern. Afterwards, a simple inhibition of return mechanism induces the model to switch attention to the second most active motion, and so on.

In addition, the wirings between the neurons within the same layer and the direction-selective attribute of some of the neurons enable our model to do a simplified constant motion tracking. If a neuron sensitive to motion direction a is activated at time t, then it passes its activation to neighboring neurons in the direction a at time t+1. In this way, the model focuses on the relevant area without recomputation of the whole motion hierarchy under the assumption that the motions do not change with time. In addition to tracking motion, a simple method for detecting the start and stop of motion is included. We applied a DOG operator to the area MST to detect motion changes [10]. Fig. 2 presents a 3D visualization of the model receiving an image sequence that contains an approaching object and a counterclockwise rotating object. Both motion patterns are correctly detected and localized.

3 Discussion and Conclusions

Due to the incorporation of functionally diverse neurons in the motion hierarchy, the output of the present model encompasses a wide variety of selectivities at different resolutions. This enables the computer simulation of the model to detect and classify various motion patterns in artificial and natural image sequences showing one or more moving objects. Most other models of biological motion perception focus on a single cortical area. For instance, the models by Simoncelli and Heeger [11] and Beardsley and Vaina [12] are biologically adequate approaches that explain some specific functionality of MT and MST neurons, respectively, but do not include the embedding hierarchy in the motion pathway. On the other hand, there are hierarchical models for the detection of motion (e.g., [13, 14]), but unlike the present model they do not provide a biologically plausible replica of the motion processing hierarchy in primates.

Another strength of our model is its mechanism of visual attention. To our knowledge, the only other motion model employing attention is the one by Grossberg, Mingolla, and Viswanathan [15], which is a motion integration and segmentation

model for motion capture. Their idea is that MST cells tuned to the winning direction have an excitatory influence on MT cells tuned to the same direction and nonspecifically inhibit all directionally tuned cells in MT. This kind of top-down influence from MST to MT has not been proved to exist yet. The current knowledge of effects of attention on single cell responses in area MT and MST suggests that cells in these areas have stronger responses when attention is directed into their RFs relative to when attention is directed outside the RF [16], which is compatible with our model.

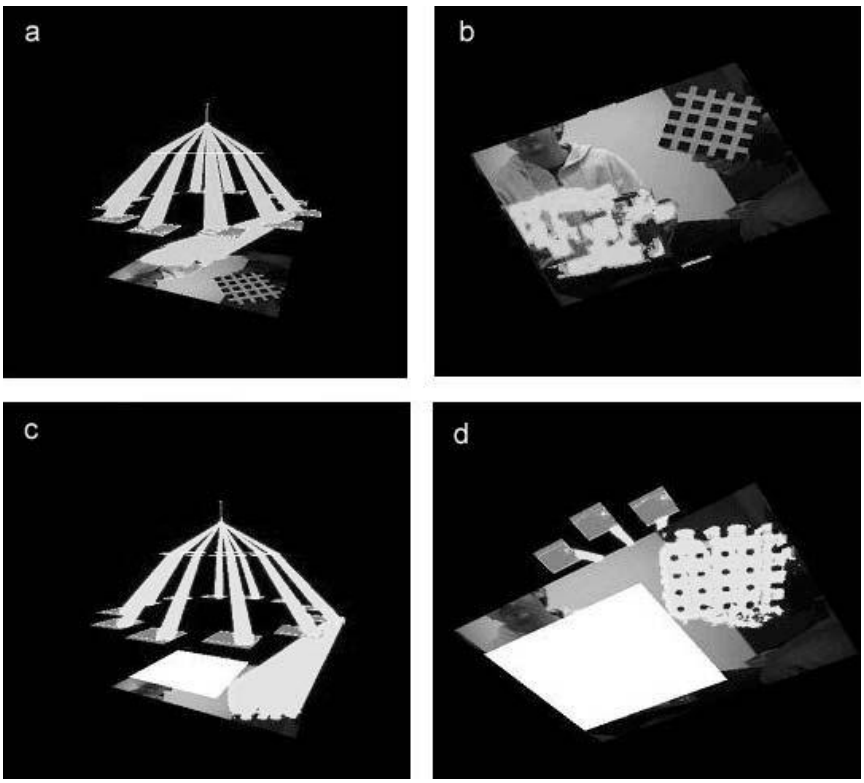


Fig. 2. Visualization of the attentional mechanism applied to an image sequence showing an approaching object and a counterclockwise rotating object at the same time. First, the model detects the approaching motion and attends to it (panel a); the localization of the approaching object can be seen most clearly from below the motion hierarchy (bright area in panel b). Then, input from the activated area is inhibited, and the model attends to the rotating motion (panels c and d).

The model has been tested on a variety of artificial and real image sequences. Simple motion patterns such as rotation, expansion, translation or combined motions with two or three patterns can be correctly recognized, localized in the image sequences and attended serially. Simple dynamic motions such as motion start, motion stop and motion pattern changes have been correctly detected as well. We

conclude that by combining four stages of motion processing with an attentional mechanism, our approach yields a biologically plausible model of visual motion processing. No current motion processing system, whether biologically inspired or not, exhibits such labeling and spatial-localization of motion patterns in image sequences.

The compatibility of our model with current neurophysiological findings and its incorporation of the diverse types of neurons found in the motion pathways provide it with predictive power for biological vision systems. Some of its predictions about activation patterns in V1, MT and MST are currently being tested in fMRI experiments on human subjects. Future work will address the perception of ego-motion, including the use of the model for controlling autonomous robots.

References

1. Orban, G.A., Kennedy, H. & Bullier, J. (1986). Velocity sensitivity and direction sensitivity of neurons in areas V1 and V2 of the monkey: Influence of eccentricity. Journal of Neurophysiology, 56 (2), 462-480.
2. Heeger, D.J. (1988). Optical flow using spatiotemporal filters. International Journal of Computer Vision, 1 (4), 279-302.
3. Lagae, L., Raiguel, S. & Orban, G.A. (1993). Speed and direction selectivity of Macaque middle temporal neurons. Journal of Neurophysiology, 69 (1), 19-39.
4. Felleman, D.J. & Kaas, J.H. (1984). Receptive field properties of neurons in middle temporal visual area (MT) of owl monkeys. Journal of Neurophysiology, 52, 488-513.
5. Treue, S. & Andersen, R.A. (1996). Neural responses to velocity gradients in macaque cortical area MT. Visual Neuroscience, 13, 797-804.
6. Graziano, M.S., Andersen, R.A. & Snowden, R.J. (1994). Tuning of MST neurons to spiral motions. Journal of Neuroscience, 14 (1), 54-67.
7. Duffy, C.J. & Wurtz, R.H. (1997). MST neurons respond to speed patterns in optic flow. Journal of Neuroscience, 17(8), 2839-2851.
8. Siegel, R.M. & Read, H.L. (1997). Analysis of optic flow in the monkey parietal area 7a. Cerebral Cortex, 7, 327-346.
9. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y., Davis, N. & Nuflo, F. (1995). Modeling visual attention via selective tuning. Artificial Intelligence, 78, 507-545.
10. Wai, W.Y.K. (1994). A computational model for detecting image changes. Master's thesis, Department of Computer Science, University of Toronto, Ontario, Canada.
11. Simoncelli, E.P. & Heeger, D.J. (1998). A model of neuronal responses in visual area MT. Vision Research, 38 (5), 743-761.
12. Beardsley, S.A. & Vaina, L.M. (1998). Computational modeling of optic flow selectivity in MSTd neurons. Network: Computation in Neural Systems, 9, 467-493.
13. Giese, M.A. (2000). Neural field model for the recognition of biological motion. Paper presented at the Second International ICSC Symposium on Neural Computation (NC 2000), Berlin, Germany.
14. Meese, T.S. & Anderson, S.J. (2002). Spiral mechanisms are required to account for summation of complex motion components. Vision Research, 42, 1073-1080.
15. Grossberg, S., Mingolla, E. & Viswanathan, L. (2001). Neural dynamics of motion integration and segmentation within and across apertures. Vision Research, 41, 2521-2553.
16. Treue, S. & Maunsell, J.H.R. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. Nature, 382, 539-541.

Real-time vision guided movement with reconfigurable Hardware

Christian Morillas, Samuel Morillas, Eduardo Ros, Antonio F. Díaz, Begoña del Pino and Francisco J. Pelayo

Departamento de Arquitectura y Tecnología de Computadores, Universidad de Granada, 18071, Granada, Spain
eduardo@atc.ugr.es, fpelayo@ugr.es

Abstract. This work summarizes the implementation and test of a vision guided mobile system. A direct sensor-motor interaction scheme leads the mobile towards the direction in which less optic flow is detected. The described system is inspired in the visuo-motor system of some insects, it uses a low-cost CMOS camera, whose digital output is captured and processed by a UP1x board of Altera. The designed digital modules that form the processing kernel of the system have been defined in VHDL. They implement real-time compression and change detection in both lateral sides of the visual field, with thresholds that are adapted depending on the global luminance of scene.

1 Introduction

The extraction and processing tasks of the visual information in real time need of high computational power. Furthermore, on one hand the visual information extraction usually requires such a computational complexity that makes difficult the use of low cost systems, but on the other hand visual information represents a very useful source for autonomous mobile systems. Clear examples of these systems are the micro-robots, in which is easy to incorporate vision front-ends through low cost micro cameras, but it is difficult to exploit this kind of sensorial information due to the low processing capabilities of these micro systems and the processing complexity required by the visual structure extraction task. For example, the visual systems of some insects such as the Dropsophila or domestic fly, extract information of the optic flow mostly driven by the ego-motion of the insect. It has been proved the existence of a very direct interaction between the sensor elements (composed eye specially sensitive movements in certain directions) and the motor elements that drive the wings. Such a direct feed-forward interaction (by means of short neuronal connection paths) provides these insects a high flying control efficiency despite their rudimentary neuronal systems [1, 2].

The hardware implementation of processing schemes based on these biological visual systems represent a valid option because simplified models [3] may be viable despite the computational resource constraints of the current implementation technologies. The current Field Programmable Logic Devices (FPLDs) are specially indicated for these kind of implementations because of their high parallelism

possibilities. This allows to allocate in the same chip the visual information extraction modules and other sensorial sources modules (multi-modal perception schemes). All this can work concurrently with the processing kernel that deal with these different information sources and makes decisions to evolve the global system in its environment.

A mobile platform (FrankeBot) [4] has been developed within the framework of a docent innovation project supported by the University of Granada (Docent Quality and Evaluation Department) that incorporates multiple sensors, digital and analog communication elements using microcontrollers and FPLDs as computational substrates. Although the system described in this work can be used in any mobile platform based on FPLDs, it was originally conceived to be integrated in the FrankeBot, in order to provide in real time measurements of the radial optic flow detected in both sides of the visual field. This optic flow is produced by the relative shift of the present features with respect to the mobile system depending on spatial and temporal differences. The optic flow provided by these features shifts in the visual field increases with higher spatial contrast patterns and when they are closer to the mobile system.

In the next section, the structure of the reference model is briefly introduced. In Section III is described the implementation of the model with diverse VHDL modules that have been synthesised with the environment Max+PlusII of Altera [5]. Finally in Section IV the final implementation is tested in a mobile platform.

2 Processing Module Structure

The flies have two composed eyes that are composed of multiple small eyes (elementary sensors) whose outputs are cooperatively collected to generate an activity pattern when a coherent movement is detected in a certain direction. A direct implementation of this movement information extraction scheme (reduced to one dimension and based on discrete analog optic sensors) is described in [2]. Furthermore, diverse VLSI approaches have been proposed that combine in the same chip, the sensors and the required analog processing circuits to extract the optic flow [3], that are called Focal-plane solutions.

In our case, the visual information is captured from a Back and White CMOS camera with digital output. In a first step the visual field is divided in two areas (left and right). Both areas will be processed separately producing different activity levels that will drive the mobile system. Each of these two areas is composed of set of elementary sensors, able to detect changes in the light intensity that reaches the receptive fields. In Fig. 1 is represented the interaction between these elementary sensors (only three in each side in this example), adding their contributions in order to produce a final estimation of the optimum direction in which the movement should evolve.

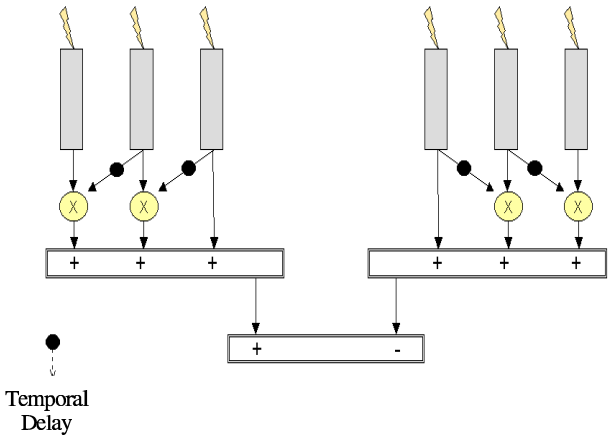


Fig. 1. Basic processing scheme to evaluate the radial optic flow. Each sensor element is a differentiator that contributes to the accumulated activity at each side of the visual field when a temporal change in the light intensity is detected.

The temporal delay and the product elements that link the neighbour sensors facilitate he contribution of those stimuli that move from the centre toward the sides and with a certain velocity. This is the movement pattern that produces an approaching object. Test results obtained from the software implementation of this model and other simplified versions have motivated the incorporation of the delay elements. The accumulated activity in one or other side of the visual field will be higher when more lateral radial flow is detected in these areas. The global accumulated activity will be calculated as the difference of these two levels (left and right sides) as illustrated in Fig. 1. This global estimation can be directly used to drive the mobile system.

In the fly the interaction between the sensors (composed eyes) and the actuators (wing motors) is almost direct; the accumulated activity is used to control the intensity that drives the wing motors. The relative movement of those efficiently controls the fly movement direction, leading to a natural tendency to get away of objects or to avoid any object with an approaching trajectory that would produce a "repelling" global optic flow signal. This natural tendency that facilitates the navigation avoiding objects is combined with an antagonist persecution and capture tendency that helps the male fly to track and reach the female fly. For this purpose the male fly eyes have a specific zone in the superior frontal eye called "love spot" [1, 2].

3 Hardware implementation of the model

As indicated in the previous section, the implementation here described uses a CMOS camera (model M4088, [6]) that uses a chip of OmniVision (OV5017, [7]).

This camera provides images of 384 columns and 288 rows (sampling the data in rows in ascending order to cover the different columns).

We make two grouping processes in order to compress the image in origin and reduce drastically the memory requirements. In a first step, we group all the pixels in the same column (adding the data). This grouping process is much easier if the scanning is done in a column order instead of a row order, and this motivates that the final orientation of the camera is rotated 90º degrees. After this reallocation, the camera provides images of 288 columns and 384 rows. In this way, the 384 data of each column are easily added (during the scanning process), obtaining what we call "macro-columns". The first consequence of this groping is that we lose any sensibility to movements in any vertical orientation, therefore we restrict our system to be able to compute only horizontal optic flow. In a second step, we group some adjacent macro-columns, computing the average, the resulting data are called "macro-pixels". In this way we gain robustness to noise although we loose resolution, we are not able to detect slight horizontal movements that could take place in this macro-pixels. Fig. 2 illustrates how each image is compressed spatially. Grouping 8 macro-columns to form a macro-pixel, we finally have 36 macro-pixels (numbered from 0 to 35) that will be used as elementary sensors for our system. The activity produced by each of this cells will be computed as the difference between absolute values calculated through the grouping procedures and the values corresponding to the previous image (temporal changes).

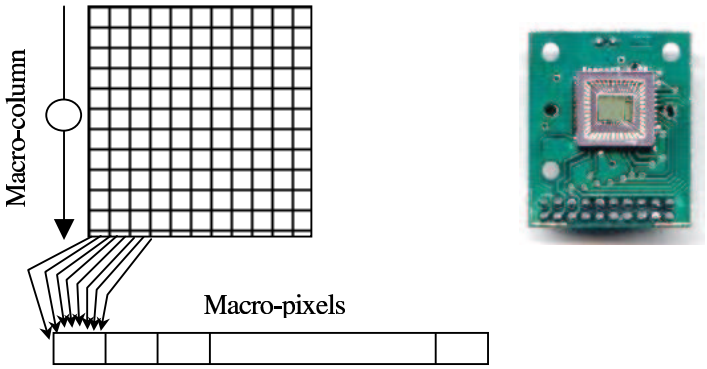


Fig. 2. Grouping of macro-columns and macro-pixels of the image captured with a rotated (90º) camera.

Now we distribute the elementary sensors to conform the left and right eye: the left eye is composed by the sensors 2 to 15 and the right eye is composed by the sensors 20 to 33. The central zone of the image (macro-pixels 16 to 19) has been eliminated as well as the lateral boundaries (macro-pixels 0,1 and 34,35).

For each frame is also obtained the average global activity that is related with the scene illumination conditions. This value is used to choose the range of significant bits in the accumulated activity, and provides the system with a certain robustness to changes in the illumination conditions (this changes are much more frequent in mobile systems than in static scenes).

The VHDL design of the system has been structured in the modules that perform the different tasks. The complete system, including the bits range selector, uses approximately 30 % of the logic cells of a CPLD-SRAM Flex-10K70 of Altera, and a 12 % of the 18Kbits of its memory blocks (EAB). The most complex module is the one that implements activity calculation stage, that uses 22 % of the logic cells of the CPLD (about 15400 logic gates). This module is structured as a three stage segmented processing pathway. This module compares the value of the captured macro-pixel with the previous value, and the result is multiplied by the delayed activity of the neighbour macro-pixel. This final magnitude is accumulated sequentially in both sides of the visual field.

4 Test of the system

For the test of the system we have used a mobile platform based on the PICBOT-2 of Microsystems Engineering [8], with an added UP-1X board of Altera with a CMOS camera. Figure 3. shows the complete system set up. An additional camera and a micro RF broadcast video to enable the remote recording of sequences from the point of view of the mobile system.

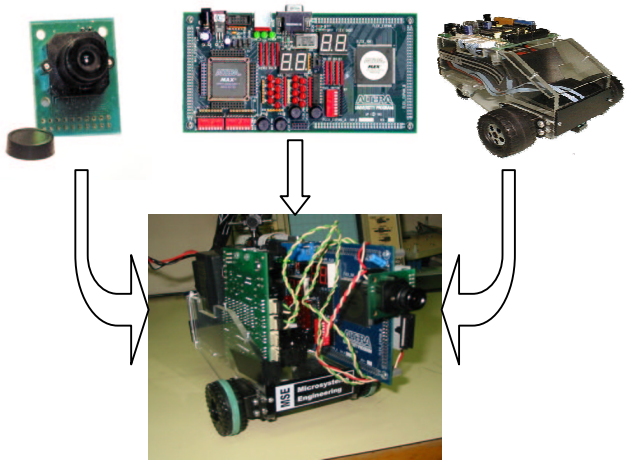


Fig. 3. Final system set up: CMOS camera, CPLD board and PICBOT-2 platform.

In Fig. 4.a can be seen the result of integration of the macro-pixels of the image (the VGA synchronism signals generation module described in [9] has been used for the visualization task). On the screen appears a dark band following the position of the black cylinder waved in front of the camera (in this case on the right side of the visual field). Finally, Fig. 4.b shows a photograph of one of the experiments of the mobile platform moving through the black cylinder wood. The response speed of the system is highly dependent of the number of processed frames per second. Further work will focus on adapting this number depending on the optic flow intensity detected in each instant.

Fig. 4. (a) Macro-pixels capturing a moving cylinder. (b) Experimental set up: Mobile robot in the cylinder wood.

Acknowledgement

This work has been carried out in the framework of the Docent Innovation Project called Hardware/Software Environment for experiments based on micro-robots, supported by the University of Granada. It has also received support from the EU research projects CORTIVIS (QLK6-CT-2001-00279, http://cortivis.umh.es) and ECOVISION (IST-2001-32114, http://www.pspc.dibe.unige.it/ecovision).

References

1. FlyBrain: *An Online Atlas and Database of the Drosophila Nervous System.* http://flybrain.neurobio.arizona.edu/
http://student.biology.arizona.edu/honors96/group11/URLS.htm
2. N. Franceschini, J.M. Pichon and C. Blanes: *From insect vision to robot vision,* Phil. Trans. Royal Society of London B 337, pp 283-294 (1992).
3. R.R. Harrison: *An analog VLSI motion sensor based on the Fly Visual System,* Ph.D. Thesis. California Institute of Technology. (1992). http://www.klab.caltech.edu/~harrison/abstracts/thesis.html
4. R. Agís, R. Carrillo, A. Cañas, B. del Pino, F.J. Pelayo: *Entorno Hardware-Software para experimentación basado en un micro-robot*. II Jornadas sobre Computación Reconfigurable y Aplicaciones, Granada, 18-20 Sept., 2002.
5. Altera. http://www.altera.com/
6. M4088 http://www.electronic-kits-and-projects.com/kit-files/cameras/d-m4088.pdf
7. OmniVision http://www.ovt.com/
8. Microsystems Engineering. http://www.microcontroladores.com/
9. Hamblen et al.: *Rapid Prototyping of Digital Systems.* Kluwer Academic Publishers. 2001.

Local Models for Dynamic Processes in Image Sequences

Hagen Spies[1,2], Tobias Dierig[2,3], and Christoph S. Garbe[3]

[1] Computer Vision Laboratory
Dept. of Electrical Engineering, Linköping University
581 83 Linköping, Sweden
hspies@isy.liu.se
[2] ICG-III: Phytosphere
Research Center Jülich, 52425 Jülich, Germany
h.spies@fz-juelich.de
[3] Interdisciplinary Center for Scientific Computing,
University of Heidelberg, INF 368, 69120 Heidelberg, Germany,
{Tobias.Dierig,Christoph.Garbe}@iwr.uni-heidelberg.de

Abstract. We present a computational framework that extends classical image velocity estimation to include more general parameters of dynamic brightness changes. The introduced method allows for an extraction of these parameters, ranging from models of linear illumination changes over diffusion and decay constants to expansion rates. We illustrate the benefit of such an extension on a real image sequence with illumination changes. We also introduce a new depth estimation technique termed depth from diffusion and apply it to some real examples.

1 Introduction

Classical image motion analysis relies on the assumption that all intensity changes are due to motion. This implies that the total derivative of the intensity g with respect to time vanishes, which is the *brightness change constraint equation* [Horn and Schunk, 1981]:

$$\frac{\mathrm{d}g}{\mathrm{d}t} = g_x u + g_y v + g_t = 0 \,. \tag{1}$$

Here we denote partial derivatives using subscripts. This concept is illustrated in Fig. 1a where the motion is along isobrightness contours. Clearly this assumption does not hold in real world situations where we encounter changes in image brightness due to variations in surface orientation or lighting conditions. An example where the intensity function also undergoes a diffusion is shown in Fig. 1b. Here the isobrightness lines will not correspond to the movement any more. The resulting velocity field computed with and without incorporation of this additional brightness change for an example image sequence is shown in Fig. 1c-f. Interestingly humans have little difficulty in perceiving the correct movement in this case.

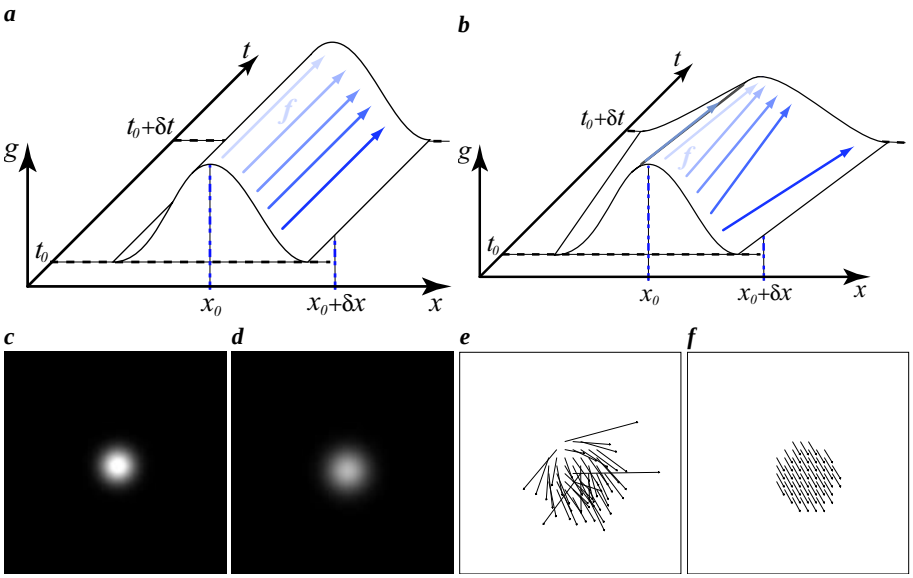


Fig. 1: *Illustration of the brightness change equation:* ***a*** *with conserved brightness,* ***b*** *with intensity changing due to diffusion.* ***c*** *first and* ***d*** *last frame of a moving Gaussian bell undergoing diffusion.* ***e*** *optical flow assuming conserved brightness and* ***f*** *estimated velocity using the extended model.*

To account for such variations the used conservation law has to be extended. Towards this end the use of multiplier and offset fields have been suggested [Negahdaripour, 1998]. Below we give a more general extension that replaces (1) by a linear partial differential equation [Haußecker et al., 1999; Haußecker and Fleet, 2001]. The novel contributions of this paper are quantitative results for a sequence with motion and illumination changes and the introduction of a new depth from X algorithm.

2 Models for Dynamic Processes

To describe more general dynamic models we allow for the intensity to vary along the trajectories we are estimating. We assume that this variation can be expressed in terms of a model function f which may depend on the intensity, time and a set of model parameters $\boldsymbol{a}$. Then the brightness change equation becomes:

$$[g_x \; g_y]\boldsymbol{v} + g_t = f(g, t, \boldsymbol{a}) \; . \tag{2}$$

Here $\boldsymbol{v}$ is the geometric velocity, for instance described by an affine motion ($\boldsymbol{v} = \boldsymbol{t} + \boldsymbol{A}\boldsymbol{x}$). The concept is very general in the sense that the parameters of any dynamic process that can be modeled by a linear partial differential equation can be quantified. Since most physical, chemical, and biological processes can be described by equations of this type, it covers many applications.

3 Total Least Squares Estimation

As all the observation data in (2) is suspect to noise it is appropriate to use a total least squares (TLS) method [Van Huffel and Vandewalle, 1991]. This method is in contrast to ordinary least squares estimation where the noise is assumed to be confined to the temporal domain. It has recently been pointed out that such a TLS model can successfully describe some observations made for the mammalian visual system [Langley, 2002].

To enable a total least squares solution we note that (2) can be written as the scalar product of a known data vector $\boldsymbol{d}$ with an unknown parameter vector $\boldsymbol{p}$: $\boldsymbol{d}^T\boldsymbol{p} = 0$. This equation poses only one constraint in the unknown parameters, thus further assumptions are needed in order to solve for the parameter field. A common smoothness requirement assumes constant parameters in a small local neighborhood of N pixel. A weighted total least squares estimate is then given by the eigenvector $\hat{\boldsymbol{e}}_n$ to the smallest eigenvalue λ_n of the so called structure tensor [Haußecker et al., 1999]:

$$\boldsymbol{J} = \boldsymbol{B} * (\, \boldsymbol{d}\, \boldsymbol{d}^T) \,, \tag{3}$$

where $\boldsymbol{B}$ is an integration kernel and $*$ denotes convolution. A good choice for $\boldsymbol{B}$ is a binomial filter as it is both symmetric and leads to a decreasing influence with distance from the considered pixel.

The above estimation is only optimal if the entries in the data vector $\boldsymbol{d}$ are uncorrelated zero mean random variables with the same noise variance [Mühlich and Mester, 1998; Van Huffel and Vandewalle, 1991]. Depending on the model used this may not be case here. To accommodate for this we simply scale the data vector accordingly, implying diagonal covariance matrices. More elaborate schemes are discussed in [Mühlich and Mester, 1999; Van Huffel and Vandewalle, 1991].

4 Experiments

In this section we demonstrate the application of the described technique to real image sequences containing illumination changes and diffusion caused by a small field of depth.

4.1 Brightness Changes

In Fig. 2a,b two frames of a sequence containing a translating plane with a random dot texture are shown. In addition to the movement the illumination changes smoothly during the sequence. The scene is illuminated via a fiber optic bundle which moves towards the scene and causes a gradual increase in intensity. Such illumination changes are easily modeled in (2) by a linear source term $f(g, t, \boldsymbol{a}) = -q$ and a translational velocity $\boldsymbol{v} = [u\ v]^T$:

$$g_x u + g_y v + g_t = -q \quad \rightarrow \quad \boldsymbol{d} = [g_x\ g_y\ 1\ g_t]^T \,; \quad \boldsymbol{p} = [u\ v\ q\ 1]^T \,. \tag{4}$$

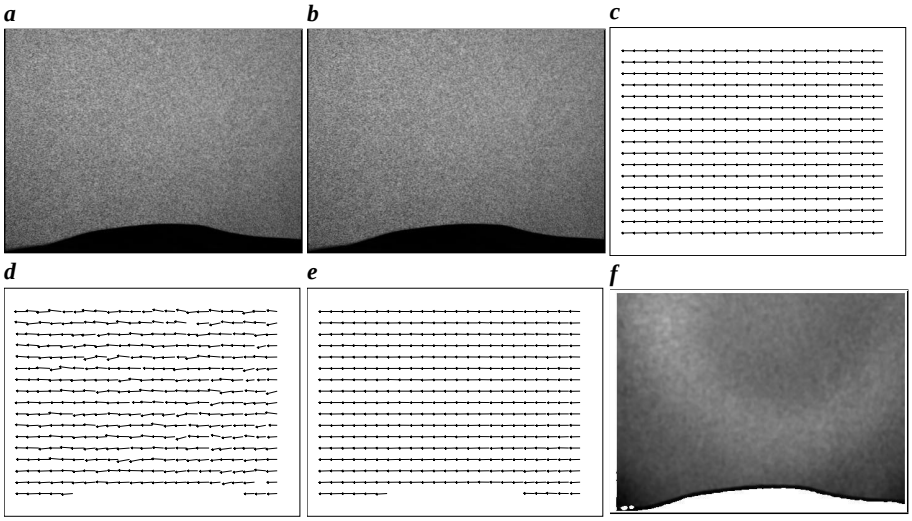


Fig. 2: *Sequence with illumination changes:* ***a*** *frame 1,* ***b*** *frame 20 and* ***c*** *correct displacement field.* ***d*** *Velocity estimated using standard optical flow constraint equation,* ***e*** *displacement when a linear source term is modeled and* ***f*** *estimated brightness changes in the range of [0, 2.5] greyvalues/frame.*

In this case there even is a constant (error free) term in the data vector. Here we simply use an error variance for this term that is two orders of magnitude smaller than that in the other terms in the scaling procedure. In practice this simplified approach usually gives good results. However, it is possible to take this error structure explicitly into account to achieve even better results [Garbe et al., 2002].

The scene consists of a plane which is moved using a linear positioner. In our laboratory setup geometric calibration information for the observing camera is available. Thus we can compute the ground truth velocity field as shown in Fig. 2c. The velocity computed assuming conserved brightness is given in Fig. 2d and that using a linear source term in Fig. 2e. In the later case we also obtain an estimate of the illumination change which is given in Fig. 2f.

Comparing the velocity fields (Fig. 2c,d,e) we can clearly see an improvement when the extended model is used. However because we do have available ground truth we can even put numbers to this improvement. The following table contains the relative error in the magnitude of the velocity, the directional error and the angular error often used in optical flow evaluations [Barron et al., 1994].

method	density [%]	rel. error [%]	dir. error [°]	ang. error [°]
standard	92.6	7.9 ± 6.3	3.3 ± 2.7	2.5 ± 1.4
extended	94.7	1.3 ± 1.2	0.5 ± 0.5	0.4 ± 0.3

Obviously there is a dramatic increase in accuracy when the illumination change is modeled.

4.2 Depth from Diffusion

An interesting application of the presented technique allows an extension of the depth from focus procedure. In its standard form a series of images with limited depth of field is acquired and at each pixel the depth is determined by the frame where it appears in focus. This technique does require telecentric lenses as the world point viewed by each pixel changes otherwise. It is common to model the blurring caused by out of focus imaging with a Gaussian point spread function. Hence the assumed underlying process is diffusion. If we thus model the changes in the intensity as a translation plus a diffusion we can capture both the motion due to the non telecentric lens and the amount of blur. Such a model can be formulated as:

$$g_x u + g_y v + g_t = -D\Delta g \quad \rightarrow \quad \boldsymbol{d} = [g_x\ g_y\ \Delta g\ g_t]^T\ ; \quad \boldsymbol{p} = [u\ v\ D\ 1]^T\ , \qquad (5)$$

where D is the diffusion constant. It can be shown that this diffusion constant is directly proportional to the distance of the observed point to the plane in focus [Dierig, 2002]. Hence D is a direct measure of depth.

In Fig. 3 two real examples are given. The displacement field is diverging as expected and the estimated depth appears to be qualitatively correct. A quantitative analysis of the recovered depth on real data has yet to be done. For a realistic setup and typical image noise we obtain a relative error in the depth below 5% on synthetic data [Dierig, 2002]. This shows that the presented general parameter estimation framework can be used successfully to compute depth from focus sequences using standard off the shelf lenses thus avoiding expensive telecentric setups and allowing for a much wider field of view.

5 Conclusion

We have presented a general framework to estimate the parameters of dynamic processes in image sequences where the assumption of conserved brightness does not hold. This has potentially a very wide application. Here we quantitatively investigated the increase in accuracy of the computed displacement field on one sequence where the illumination changes. Furthermore we introduced a novel algorithm termed *depth from diffusion* to compute depth from focus series taken with non telecentric cameras. This is achieved by modeling blur as a diffusion process.

Acknowledgements. Part of this work has been founded under the DFG research unit "Image Sequence Analysis to Investigate Dynamic Processes" (FOR240) and by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD).

Bibliography

J. L. Barron, D. J. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.

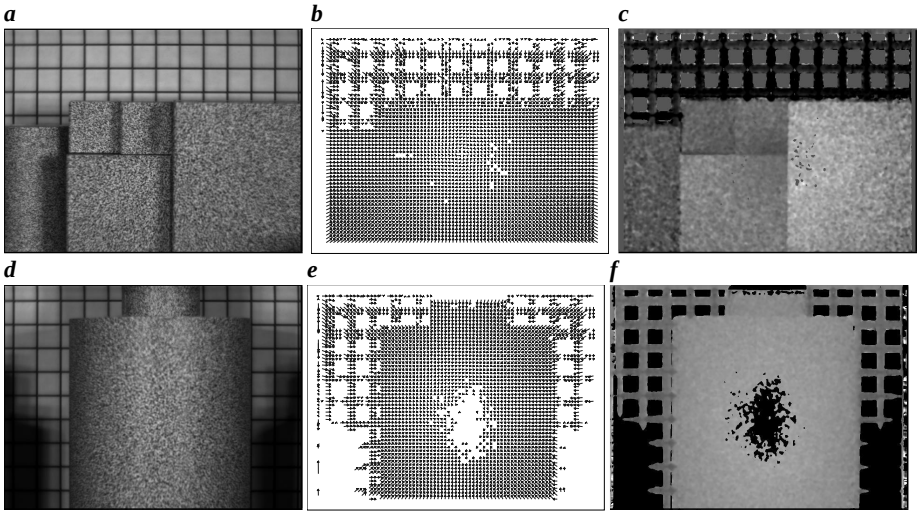


Fig. 3: *Depth from Diffusion. First example:* ***a*** *one image of the sequence,* ***b*** *displacement field and* ***c*** *estimated depth. Second example:* ***d*** *original image,* ***e*** *displacement field and* ***f*** *estimated depth.*

T. Dierig. *Gewinnung von Tiefenkarten aus Fokusserien*. PhD thesis, University of Heidelberg, Heidelberg, Germany, July 2002.

C. Garbe, H. Spies, and B. Jähne. Mixed ols-tls for the estimation of dynamic processes with a linear source term. In *DAGM*, Lecture Notes in Computer Science, Zürich, Switzerland, September 2002. Springer.

H. Haußecker and D. J. Fleet. Computing optical flow with physical models of brightness variation. *PAMI*, 23(6):661–673, June 2001.

H. Haußecker, C. Garbe, H. Spies, and B. Jähne. A total least squares framework for low-level analysis of dynamic scenes and processes. In *DAGM*, pages 240–249, Bonn, Germany, 1999. Springer.

B. K. P. Horn and B. Schunk. Determining optical flow. *Artificial Intelligence*, 17: 185–204, 1981.

K. Langley. Motion perception and motion estimation by total-least squares. *Spatial Vision*, 15(2):171–190, 2002.

M. Mühlich and R. Mester. The role of total least squares in motion analysis. In *ECCV*, pages 305–321, Freiburg, Germany, 1998.

M. Mühlich and R. Mester. Subspace methods and equilibration in computer vision. Technical Report XP-TR-C-21, Institute for Applied Physics, Goethe-Universität, Frankfurt, Germany, November 1999.

S. Negahdaripour. Revised definition of optical flow: Integration of radiometric and geometric cues for dynamic scene analysis. *PAMI*, 20(9):961–979, September 1998.

S. Van Huffel and J. Vandewalle. *The Total Least Squares Problem: Computational Aspects and Analysis*. Society for Industrial and Applied Mathematics, Philadelphia, 1991.

Dynamic visual scenes

Drawing an Illusion across Primary Visual Cortex: Line-Motion revealed by Voltage-Sensitive Dye Imaging

Dirk Jancke[1,2], Frédéric Chavane[1], Amos Arieli[1] & Amiram Grinvald[1]

[1] Department of Neurobiology and the Grodetzky Center for Studies of Higher Brain Function, Weizmann Institute of Science, Rehovot 76100, Israel.

[2] Allgemeine Zoologie und Neurobiologie, Ruhr-University Bochum, 44780, Germany.

e-mail: jancke@neurobiologie.ruhr-uni-bochum.de

Abstract. Visual illusions reveal fundamental processing mechanisms of which we are unaware during our daily perceptual experiences. The "line-motion" illusion[1,2] consists of a flashed dot followed by a flanking bar with some time delay. Instead of sensing the bar at once, subjects report an illusory line drawing, away from the dot (see Fig. 1). Using voltage-sensitive dye optical imaging, we visualized line-motion in real-time on the surface of cat area 18.

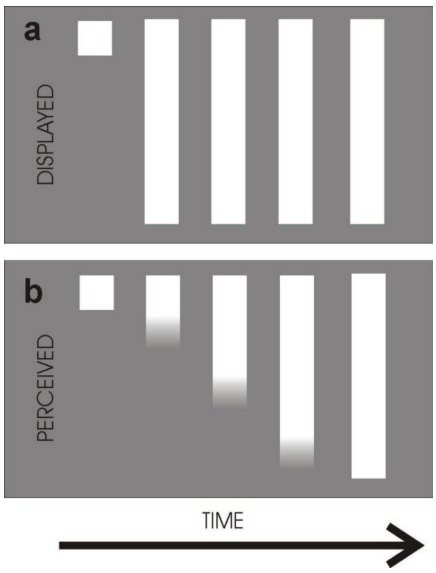


Fig. 1. The Line-Motion Illusion. a) A square of light ("pre-cue") is presented just before a bar stimulus. b) Instead of sensing the bar at once, subjects report an illusory line drawing starting from the pre-cued location.

Wertheimer (1921)[3] and Kenkel (1913)[4] made the surprising observation that even stationary stimuli can give the impression of motion, the "gamma movement". An appearing local stimulus is perceived as expanding or otherwise as contracting when it disappears from a homogenous background. This effect can be polarized and strengthened if a local cue is presented adjacent to an elongated bar stimulus. In such a case, illusory motion is seen away from the cue[5].

The line-motion illusion was attributed to an attentional gradient that facilitates processing in the surround of the pre-cueing dot. Although many alternative explanations exist, most psychophysicists encircled the origin of the line-motion illusion in early processing stages likely after binocular fusion. Yet, in need of a neurophysiological method that offers both high temporal and high spatial resolution, it remained unclear which neural mechanisms could account for building-up motion within a bar.

In order to visualize cortical line-motion we used optical imaging of voltage-sensitive dyes in area 18 of the anaesthetized and paralyzed cat[6,7] This technique measures changes in synaptic potentials of neural populations, thus monitoring evoked activity in real-time across a certain cortical region that entirely represents the stimuli shown[8].

The spatio-temporal characteristics of activity evoked by a flashed square alone can be described in two steps: 1.) Stimulus appearance evokes "subthreshold" propagating activity that gradually slows down as the response amplitude increases. 2.) Only at high levels activity stays local, i.e. motionless. The deceleration of propagating activity could be the result of a filter process that transmits activity through horizontal axons onto the wide arborisation of neural dendrites. How does a flashed square then affect the response to a subsequently presented bar?

In the line-motion condition, the "subthreshold" propagating activity is rapidly enhanced (15 ms after the bar onset) by the following bar and thus, expressed above threshold at a speed guided by the spatio-temporal properties in response to the flashed square alone. This leads to a very significant wave front that moves at a constant speed, away from the pre-cued location. As a result, the cortical surface is representing the progressive line drawing illusion.

Our results are in line with studies that referred to the phenomenon as motion induction by pre-attentive facilitation or as an apparent-motion process with no need of attention per se. However, high-level processes might modulate speed and shape of propagating activity. There are evidences for attention-related components, operating on a slower time scale on the perception of the line-motion illusion. It has also been shown that line-motion can be induced voluntary. Thus, in the behaving subject, additional mechanisms are interacting along the visual pathway. We suggest that the cortical representation of the line-motion illusion uncovers an "automatic" process in primary visual cortex that may serve to compute motion at higher processing stages and guide bottom-up attention.

Bringing together psychophysics and neurophysiology using awake animals in future studies may reveal influences of top-down attention and stimulus attributes on the representation of speed in primary visual cortex.

Acknowledgements

We thank Rina Hildesheim for synthesizing the dye, RH-1692, Dov Ettner, Yuval Toledo and Chaipi Wijnbergen for technical assistance, Anirudh Gupta, Dahlia Sharon, Eyal Seidemann, Hamutal Slovin and Ivo Vanzetta for important comments and discussions. This work was supported by MINERVA Foundation, Germany (D.J.), and Marie Curie E.-U. Fellowship (F.C.) and by grants from the Grodetsky Center, the Goldsmith and the Korber and ISF Foundations (A.G).

References

[1] Hikosaka O., Miyauchi, S. & Shimojo, S. Focal visual attention produces motion sensation in lines. *Investigative Ophthalmology and Visual Science*, **32**, 716 (1991).

[2] Hikosaka O., Miyauchi, S. & Shimojo, S. Focal visual attention produces illusory temporal order and motion sensation. *Vision Res* **33**, 1219-1240 (1993).

[3] Wertheimer, M. Experimentelle Studien über das Sehen von Bewegung. *Zeitschrift für Psychologie* **61**, 162-265 (1912).

[4] Kenkel F. Untersuchungen über den Zusammenhang zwischen Erscheinungsgröße und Erscheinungsbewegung bei einigen sogenannten optischen Täuschungen. *Zeitschrift für Psychologie* **67**, 358-449 (1913).

[5] Kanizsa, G. Sulla polarizzazione del movimento gamma. *Archivio di Psicologia, Neurologia e Psichiatria* **3**, 224-267 (1951).

[6] Grinvald, A., Anglister, L., Freeman, J,A,, Hildesheim, R. & Manker, A. Real-time optical imaging of naturally evoked electrical activity in intact frog brain. *Nature* **308**, 848-850 (1984).

[7] Grinvald, A., Lieke, E., Frostig, R. & Hildesheim, R. Cortical point-spread function and long-range lateral interactions revealed by real-time optical imaging of macaque monkey primary visual cortex. *J Neurosci* **14**, 2545-2568 (1994).

[8] Shoham, D., Glaser, D.E., Arieli, A., Kenet, T., Wijnbergen, C., Toledo, Y., Hildesheim, R. & Grinvald, A. Imaging cortical dynamics at high spatial and temporal resolution with novel blue voltage-sensitive dyes. *Neuron* **24**, 791-802 (1999).

Object Representation Through Transient Neural Dynamics

Udo A. Ernst[1], Axel Etzold[1], Michael H. Herzog[2], and Christian W. Eurich[1]

[1] Institute for Theoretical Physics, Otto-Hahn-Allee 1, University of Bremen, Postbox 330 440, D-28334 Bremen, Germany, {uernst,aetzold,eurich}@physik.uni-bremen.de

[2] Institute for Human Neurobiology, Argonnenstr. 3, University of Bremen, D-28209 Bremen, Germany, michael.herzog@uni-bremen.de

Abstract. To survive in a complex and ever changing environment, an organism has to cope with sensory stimuli often varying on a short time scale. Signal processing in the nervous system should, therefore, be dynamical and fast: often it is not feasible to wait until the neural activation pattern of the brain settles into a steady state before an appropriate reaction is initiated. Here, we study visual processing at the brink of its temporal and spatial resolution by using the recently discovered shine-through effect. We show how transient perception can arise by neural dynamics described by a Wilson-Cowan type neural network. Moreover, our results impose restrictions on the time and length scales involved in visual cortical processing, and allow to predict under which conditions a masked stimulus reaches visibility.

1 Introduction

One of the fundamental questions in visual processing is how a complex, time-varying stimulus is segmented and interpreted by the neural hardware to form a coherent percept. A particularly useful strategy to tackle this question is to study the *limitations* of this process – because those limits effectively restrict the search for possible mechanisms behind the information processing going on in the brain.

Here, we present a new psychophysical effect, shine-through, that allows to investigate the dynamics of transient perception in great temporal and spatial detail. In contrast to many other pyschophysical studies, the stimuli as well as the percepts are non-static, and therefore yield valuable conclusions about the time-course of visual signal processing. These dynamical phenomena and their underlying mechanisms are studied in a neural network model, where we focus explicitly on the *transients*, and not on the fixed points or limit cycles that are normally investigated.

2 Shine-through

In the shine-through effect a target element, for example a *vernier* (two abutting lines with displacement d), precedes a homogeneous and extended grating

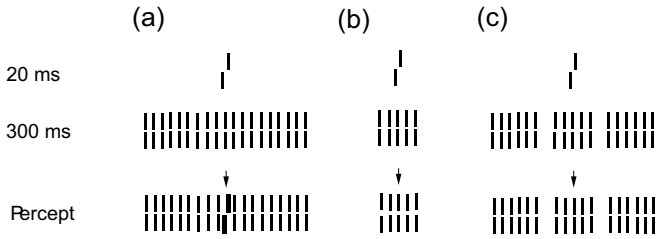


Fig. 1. A vernier presented for 20 ms precedes a grating of various spatial layout presented for 300 ms. (a) Only for a homogeneous grating shine-through occurs: the vernier appears as a *transient*, short flash superimposed on the grating looking wider, brighter, and even longer than the vernier really is. For (b) and (c), the vernier element is rendered invisible by the masking gratings – no shine-through occurs.

(Fig. 1(a)) displayed for 300 ms [1, 2]. In spite of the masking grating, for trained observers the vernier is clearly visible even if display times are as short as 20 ms, i.e. in the range of a few neural spikes. Visibility is assessed as the threshold displacement d of the vernier necessary to yield 75% correct discrimination performance.

Shine-through diminishes dramatically if the grating comprises less than seven elements (Fig. 1(b)). Shine-through ceases also for spatially inhomogeneous gratings. For example, a grating containing gaps renders the vernier completely invisible (Fig. 1(c)). From a figure-ground-segmentation point of view, the grating is parsed into three independent entities. Performance deteriorates since the central part is a small grating not allowing shine-through (see Fig. 1(b)).

In all three conditions the target vernier either appears as a *transient* entity or is rendered invisible by changes of the spatio-temporal layout of the masking grating. Hence, the underlying mechanisms point to a system in which neurons compete with each other. Although the psychophysical results suggest that high level Gestalt factors cause the changes in perception and performance, we show in the following that a simple model can account for the empirical findings – without including any explicit high order Gestalt processing.

3 Model

Our model employs the horizontal axis x of the visual field only and neglects the vertical spatial direction and the orientation tuning of cortical visual cells to simplify analysis. The network (Fig. 2) consists of a one-dimensional layer with one excitatory and one inhibitory neuronal population, mutually connected with coupling kernels $W_{\{e,i\}}$, with typical length scales $\sigma_{\{e,i\}}$,

$$W_{\{e,i\}}(x - x') = \frac{1}{\sqrt{2\pi\sigma^2_{\{e,i\}}}} \exp\left(-\frac{(x - x')^2}{2\sigma^2_{\{e,i\}}}\right) . \tag{1}$$

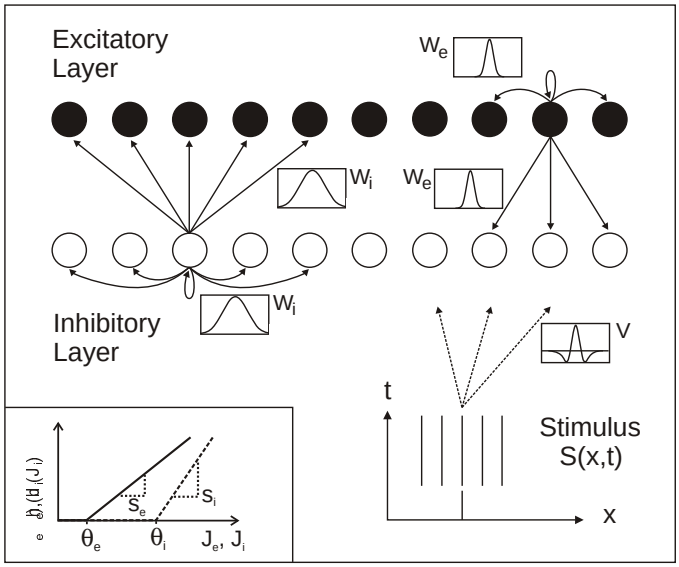


Fig. 2. Structure of model employed in the simulations. A spatio-temporal stimulus $S(x,t)$ is filtered by a difference of Gaussians and projected onto two populations in a one-dimensional neuronal layer. The two populations, an excitatory and an inhibitory one, are mutually coupled with synaptic weight functions described by the Gaussian kernels W_e and W_i, respectively. The inset shows the neuronal gain functions mapping the synaptic inputs $J_{\{e,i\}}$ to the firing rates $h_{\{e,i\}}$.

The dynamics of the system are given by a set of Wilson-Cowan type equations [3] (for an overview see [4]) for the excitatory activities A_e and inhibitory activites A_i of the populations,

$$\begin{aligned}\tau_e \frac{\partial A_e(x,t)}{\partial t} &= -A_e(x,t) + h_e\left\{w_{ee}\left(A_e \star W_e\right)(x,t) + \right. \\ &\quad \left. + w_{ie}\left(A_i \star W_i\right)(x,t) + I(x,t)\right\} \qquad (2)\\ \tau_i \frac{\partial A_i(x,t)}{\partial t} &= -A_i(x,t) + h_i\left\{w_{ei}\left(A_e \star W_e\right)(x,t) + \right. \\ &\quad \left. + w_{ii}\left(A_i \star W_i\right)(x,t) + I(x,t)\right\}\,, \qquad (3)\end{aligned}$$

with $w_{ee}, w_{ei}, w_{ie}, w_{ii}$ denoting coupling strengths, $\tau_{\{e,i\}}$ denoting time constants, $I(x,t)$ denoting the efferent input, and $h_{\{e,i\}}$ describing the gain functions (see Fig. 2 inset). The stars in Eqs. (2)-(3) denote convolutions of the population actitivies with the coupling functions as e.g. for

$$w_{ee}\left(A_e \star W_e\right)(x,t) = w_{ee} \int_{-\infty}^{\infty} A_e(x',t) W_e(x-x')\, dx'\,. \qquad (4)$$

The convolution of the efferent coupling kernel V,

$$V(x-x') = \frac{1}{\sqrt{2\pi\sigma_E^2}} \exp\left(-\frac{(x-x')^2}{2\sigma_E^2}\right) - \frac{1}{\sqrt{2\pi\sigma_I^2}} \exp\left(-\frac{(x-x')^2}{2\sigma_I^2}\right), \quad (5)$$

with the spatio-temporal pattern $S(x,t)$ modeling the stimulus sequences used in the experiment (see Fig. 1), yields the efferent input I converging onto both populations, $I(x,t) = (S \star V)(x,t)$. For the mutual couplings defined in Eq. (1), the range of inhibition is chosen to be larger than the range of excitation; also, we assumed that recurrent input dominates over efferent input [5, 6].

The psychophysical detection threshold d was related to the activity profiles coming out of the model via the time interval T the excitatory activity $A_e(0,t)$ in the center population remained above an observation threshold h_t. While the exact relationship between these two measures is analyzed elsewhere [7], let us note here that with a longer duration T, the more information about the vernier's displacement can be gathered, and a smaller detection threshold d can therefore be expected.

4 Results

In the shine-through effect, the vernier appears as a bright flash superimposed on the grating. Therefore, the processing of the vernier signal is expected to occur as a *transient* in the neural dynamics and not as a steady state. Numerical results for the stimulus conditions (a)-(c) of Fig. 1 are shown in Figs. 3(a)-(c). The color-coded activities of the excitatory populations show peaks at the position of the vernier and at the edges of the gratings, whereas almost no activity emerges for the inner grating elements. The time course of the activity of the central neural population in Figs. 3(a)-(c) is shown in Fig. 3(d). The central peak in the condition with the small grating (Figs. 1(b),3(b)) decays faster as compared to the condition with the extended grating (Figs. 1(a),3(a)). This behavior is explained by the strong inhibition radiating from the active neurons representing the nearby edges of the grating comprised of only 5 elements (see arrow in Fig. 3(b)). However, if the extended grating comprises 25 elements, the edges are too remote to exert a substantial inhibitory influence on the center (Fig. 3(a)). Thus, the activity elicited by the vernier is sustained by feedback excitation, and decays much more slowly than in condition (c). Inserting gaps in the grating of 25 elements (see Fig. 1(c)) introduces inhomogeneities leading to an enhanced activation at these gaps whose inhibitory surrounds suppress the vernier activity as fast as in the 5 element condition (see arrow in Fig. 3(c)).

Perceptually, the fast suppression of the vernier activity by the small central grating shown in Figs. 1(b) and (c) leads to a complete masking of the vernier element. On the other hand, conditions which allow a longer persistence of the vernier activity like the one in Fig. 1(a) result in a conscious perception of the vernier and its displacement. Thus, the occurrence of shine-through can be explained with the transient dynamics of a Wilson-and-Cowan type model.

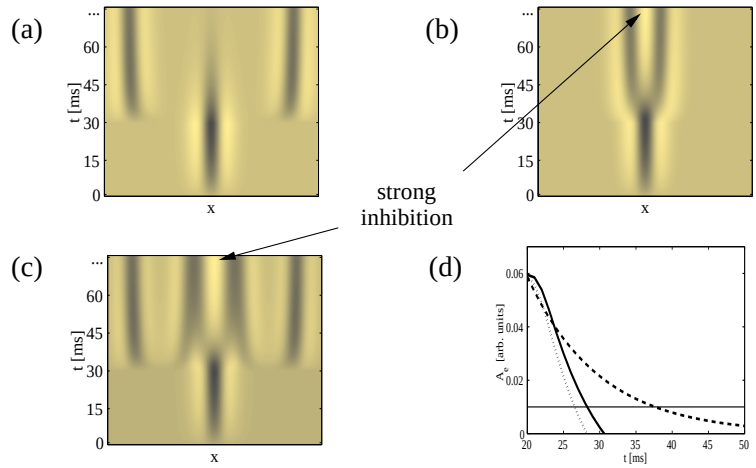


Fig. 3. Spatio-temporal activation patterns emerging from the Wilson-Cowan model for the three different masking conditions in Fig.1. The activation levels of the excitatory population are color-coded (dark for high activation). The ordinates correspond to the location of the neuronal population x, and time t in milliseconds is shown on the abscissa. (a) Vernier activity persists since peaks of neural activity appear only at the distant edges of the 25-element grating, exerting no inhibition on the activity corresponding to the vernier. In (b) and (c), the activities corresponding to the edges of the 5-element grating rapidly suppress vernier activity. The time course of the activation of the center population is shown in (d), where the solid and dotted curves correspond to the conditions modelled in (b) and (c), respectively, while the dashed curve shows the slower decay from the condition modelled in (a). The thin line in (d) shows the observation threshold h_t choosen to be $h_t = 0.008$.

5 Summary and Discussion

Our results demonstrate that a structurally simple model based on only two partial differential equations is sufficient to explain psychophysical phenomena of the visibility of masked stimuli. Transient activation of a neuronal population instead of fixed points of its dynamics determines the visibility of the target element. Moreover, global, Gestalt-like perceptual conditions can be explained through simple interactions in topologically arranged neural layers.

The mechanisms behind the observed model dynamics can be summarized in terms of the most important model parameters: The convolution of the stimulus with the Mexican-hat efferent coupling kernel having the length scales $\sigma_{E,I}$, yields enhanced input at the edges of a regularly spaced grating, while input from the inner elements is suppressed. The activity subsequently emerging at the aforesaid edges then suppresses any activity in a distance of $\sigma_i \approx 3\ d_{bar}$, being the length scale of the recurrent inhibition. This "edge detection" on a length scale d_{bar} [8], and the "competition" between activity on a length scale

of 3 d_{bar}, leads to the differences between the shine-through (Fig. 3(a)), and the other two stimulus conditions (Figs. 3(b) and (c)). These differences are most pronounced when the ratio of the excitatory and inhibitory time constants, τ_e/τ_i, gets large.

When interpreting the experiments and simulations in terms of a figure-ground segmentation process, one may draw the following conclusions from the observed dynamics. First, segmentation enhances inhomogeneities in a stimulus being presented – in our case, the inhomogeneities correspond to the edges of the masking gratings. Second, segmentation is a time-consuming process: the vernier activity has to be high enough, and has to persist for a sufficiently long time, in order to be perceived correctly. This condition is fulfilled only in Fig. 1(a), while in Figs. 1(b) and (c), the segmentation of the masking grating rapidly disrupts the segmentation of the vernier. Third, in contrast to the previous conclusion, the *onset* of segmentation is very fast – even slight temporal and spatial changes to the shine-through condition Fig. 1(a) render the vernier invisible (data not shown, [7]). And finally, two on-going segmentation processes do not interfere when the features of the stimuli are well separated either in time *or* in space (see Figs. 1(a) and 3(a)).

Supported by the Sonderforschungsbereich 517 "Neurocognition" (M.H.H., C.W.E., and U.A.E.) and the Volkswagen Stiftung, Project 5425 (U.A.E.).

References

1. M. H. Herzog and C. Koch. Seeing properties of an invisible object: feature inheritance and shine-through. *PNAS*, 98:4271–4275, 2001.
2. M. H. Herzog, M. Fahle, and C. Koch. Spatial aspects of object formation revealed by a new illusion, shine-through. *Vision Research*, 41:2325–2335 and Erratum in Vision Research 42 (2001) 271, 2001.
3. H. R. Wilson and J. D. Cowan. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13:55–80, 1973.
4. U.A. Ernst and C.W. Eurich. Cortical population dynamics and psychophysics. In M.A. Arbib, editor, *2nd Handbook of Brain Theory and Neural Networks*. MIT Press, Boston MA, in press.
5. R. Ben-Yishai, R. Bar-Or, and H. Sompolinsky. Theory of orientation tuning in visual cortex. *PNAS*, 92:3844–3848, 1995.
6. U.A. Ernst, K.R. Pawelzik, C. Sahar-Pikielny, and M.V. Tsodyks. Intracortical origin of visual maps. *Nature Neurosci.*, 4:431–436, 2001.
7. M.A. Herzog, U.A. Ernst, A. Etzold, and C.W. Eurich. Local interactions in neural networks explain global effects in the masking of visual stimuli. *submitted to Neural Computation*, 2002.
8. Z. Li. Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex. *Network: Comput. Neural Syst.*, 10:187–212, 1999.

Saccadic Undershoots and the Relative Localization of Stimuli[1]

Sonja Stork & Jochen Müsseler
Max Planck Institute for Psychological Research
Cognition and Action, Amalienstr. 33, D-80799 Munich
Email: stork@mpipf-muenchen.mpg.de
http://www.mpipf-muenchen.mpg.de/~stork

Abstract. When observers are asked to localize the peripheral position of a probe with respect to the mid-position of a spatially extended comparison stimulus, they tend to judge the probe as being more peripheral than the mid-position of the comparison stimulus. We investigated the relationship between this perceived mislocalization and saccadic undershoots. The findings show that the mislocalization corresponds to the saccadic behavior. Moreover, differences in saccadic undershoots to the stimuli can be used to estimate quantitatively the amount of relative mislocalization.

1 Introduction: A Relative Mislocalization

Visual localization acuity measured with long-presented stationary stimuli is of high precision. However, several studies indicated that spatial acuity is considerably poorer under less optimal viewing conditions. We studied the ability to localize a flashed stimulus and its relationship to saccadic eye movements with a relative judgment task (cf. Fig. 1). When observers judge the peripheral position of a probe with respect to the mid-position of a spatially extended comparison stimulus, the probe is seen more peripheral than the mid-position of the comparison stimulus [4]. We suggested and found evidence that this relative mislocalization emerges from different absolute mislocalizations. When observers point to the position of the spatially extended comparison stimulus they tend to localize it more foveally than the spatially less extended probe (see also [7]).

Comparable foveal tendencies in absolute localizations are known from eye-movement behavior. Eyes tend to undershoot a peripherally presented target, before they reach it with a corrective saccade [1]. Additionally, this undershoot seems to be more pronounced with a spatially extended stimulus [2]. If these results based on a spatial map, which is used by both the perceptual judgment task and the saccade task, the probe's relative position should be perceived more peripheral when compared with the mid-position of the comparison stimulus.

However, the mislocalization is only observed when stimuli are flashed successively (i.e., typically with a stimulus onset asynchrony of about 120 ms). In this case two configurations with different spatial information have to be superimposed and the relative mislocalization between stimuli can emerge. In contrast, when stimuli are flashed simultaneously, they can be processed in one spatial map. Accordingly, the

[1] This research was funded by the German Science Foundation (AS 79/3).

localization judgment of the probe relative to the comparison stimulus was found to be more or less precise with simultaneous presentation [4].

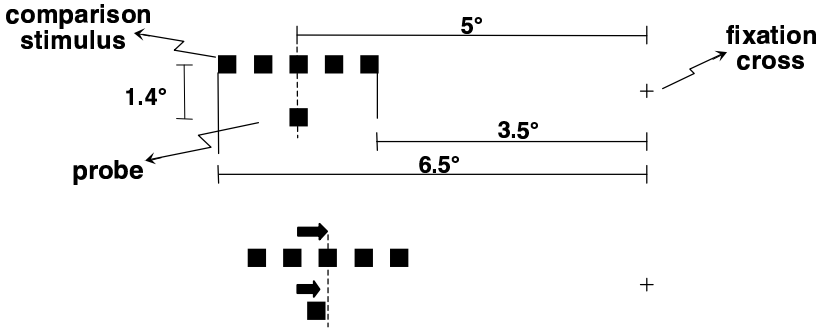


Fig. 1. Stimulus presentation (upper panel) and stimulus perception (lower panel). The perceived mislocalization of the probe relative to the mid-position of the comparison stimulus (lower panel) is assumed to emerge from different absolute localizations (as indicated by the arrows) of the stimuli with respect to the fixation cross.

To conclude our preliminary interpretation of the mislocalization is based on the assumption that saccadic tendencies contribute to the position codes of a spatial map. This map is used to determine the perceived localizations [4, 8]. In contrast, other accounts suggest that eye movements are specified in a direct manner independent of the perceived representation [3]. Several phenomena demonstrate dissociations between perception and action indicating different neural pathways for goal-directed behavior and for the perception of objects. Accordingly, the dorsal pathway is assumed to be involved in the execution of saccades (especially in the medially parieto-occipital sulcus, V5), while the ventral pathway is assumed to be involved in visual illusions. If this is correct, saccadic behavior need not match with the mislocalization observed in the relative judgment task. In order to clarify this issue, we examined whether and how saccadic undershoots are related to the relative judgments (for details see [6]).

2 New Findings and Conclusions

In Experiment 1, saccades to the comparison stimulus or the probe were compared with the perceptual judgments. In the saccade task, subjects were instructed to execute a saccade to a target (the probe or the mid-position of the comparison stimulus) as fast as possible. In the judgment task the position of the probe was varied with respect to the mid-position of the comparison stimulus and subjects were asked which stimulus was more peripheral – the upper one or the lower one?

If the saccadic behavior and perceptual judgment correspond, saccades to the comparison stimulus should show a stronger undershoot than to the probe. Indeed, results show that observers produce smaller saccadic amplitudes to the comparison stimulus than to the probe. This effect in saccades was observed when stimuli were

presented separately, that is, only the probe or the comparison stimulus appeared on the screen.

The subsequent experiments were run in order to check whether the eccentricity of stimuli presentation exert an influence on both the judgments and the saccades. As in previous experiments [4], the perceived relative mislocalization increased with eccentricity. In contrast, the saccadic undershoot did not show a corresponding effect, when both stimuli were presented separately (Experiment 2). However, they corresponded to the perceived relative mislocalizations when both stimuli appeared on the screen (as in the relative judgment task). In this case, subjects' task was to generate a saccade to the probe or the mid-position of the comparison stimulus and to ignore the other stimulus (Experiment 3). The finding that only in this case saccadic behavior and perceptual judgment correspond demonstrate the importance of targets' context on the saccadic behavior.

In sum, the pattern of results indicates that – if comparable temporal and spatial configurations are used – the saccadic behavior corresponds qualitatively with the perceived relative mislocalization. In an additional analysis the relationship between both measures was analyzed quantitatively. In a first step of this analysis, the outer edge of the stimuli and further stimulus parameters of the present experiments proved to be important variables to determine the saccadic landing positions (for details see [5]). In a second step, these landing positions were used to estimate the relative mislocalization by computing the difference between the landing positions to the probe and the comparison stimulus.

Observed and estimated relative mislocalizations of the present experiments and of a previous study [4] are plotted in Figure 2. On the one hand, the plot shows a high positive correlation. Thus, it is possible to estimate the perceived relative mislocalization by the variables determining the saccadic behavior. On the other hand, the slope and the intercept of the regression line does not equal 1 and 0, respectively. Accordingly, one could still claim a dissociation between saccadic behavior and perceptual judgment.

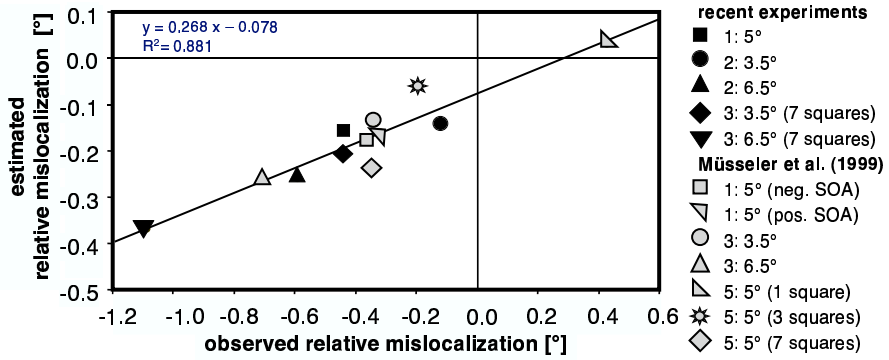


Fig. 2. Observed and estimated relative mislocalization.

Nevertheless, our findings demonstrate an obvious association between both measures. An interesting problem to think about is why eye movements undershoot the target at all and – more critically – why the system does not adapt to this error. One might speculate that the undershoot is an inherent property of any motor system, probably because it is easier to correct a movement in its direction than in the opposite direction. Another argument would be that with an undershoot the retinal image of the target remains in the same cortical hemifield and the system need not switch to the other hemifield. A last, but not least possibility comes from considering more ecological conditions. Usually, targets do not enter the visual field instantaneously but appear in the visual field and move into it. It could be a saccadic undershoot anticipates this movement.

Our interpretation is in accordance with the assumption that the saccadic behavior together with sensory information establishes perceived space. In other words, we assume that the system in charge of the guidance of eye movements is also the system that provides the metric of perceived visual space [8]. The position code for the localization judgment and for saccades shows comparable tendencies, indicating a common mechanism for both purposes. However, the differences in estimated landing positions of the eyes were less pronounced than the relative observed mislocalizations indicating a late modulation of the perceptual judgment.

References

[1] Deubel, H., Wolf, W., & Hauske, G. (1982). Corrective saccades: Effect of shifting the saccade goal. Vision Research, 22, 353-364.

[2] Findlay, J. M., Brogan, D., & Wenban-Smith, M. G. (1993). The spatial signal for saccadic eye movements emphasizes visual boundaries. Perception & Psychophysics, 53, 633-641.

[3] Milner, A.D., & Goodale, M. A. (1995). The visual brain in action. Oxford, UK: Oxford University Press.

[4] Müsseler, J., van der Heijden, A. H. C., Mahmud, S. H., Deubel, H., & Ertsey, S. (1999). Relative mislocation of briefly presented stimuli in the retinal periphery. Perception & Psychophysics, 61, 1646-1661.

[5] Stork, S. (2002). Blickbewegungen und die Lokalisation von stationären und bewegten Reizen [Eye movements and the localization of stationary and moving stimuli]. Unpublished doctoral dissertation. Ludwig-Maximilians University, Munich.

[6] Stork, S., Müsseler, J., & van der Heijden, A. H. C. (submitted). Saccadic eye movements and a relative mislocalizations with briefly presented stimuli

[7] van der Heijden, A. H. C., van der Geest, J. N., de Leeuw, F., Krikke, K., & Müsseler, J. (1999). Sources of position-perception error for small isolated targets. Psychological Research, 62, 20-35.

[8] van der Heijden, A. H. C., Müsseler, J., & Bridgeman, B. (1999). On the perception of positions. In G. Aschersleben, T. Bachmann, & J. Müsseler (Eds.). Cognitive contributions to the perception of spatial and temporal events (pp. 19-37). Amsterdam: Elsevier.

Cognitive influences on visual processing

How does the ventral pathway contribute to spatial attention and the planning of eye movements?

Fred H. Hamker

California Institute of Technology, Division of Biology 139-74,
Pasadena, CA 91125, USA
fred@klab.caltech.edu
http://www.klab.caltech/∼fred.html

Abstract. Cortical organization of vision appears to be divided into two pathways: the ventral pathway and the dorsal pathway. Models of vision have generally adopted this separation into a functional division such that recognition is supposed to be located in the ventral pathway and spatial attributes are processed in the dorsal pathway. I suggest a less distinct separation. According to my model the ventral pathway contributes to the selection of the location of an object by feedback connections. Those projections localize the object of interest by transferring information about its features in IT to cells with smaller receptive fields in V4 and earlier. I demonstrate the performance of the model in a visual search task which demands an eye movement towards a target.

1 Introduction

Visual perception is proposed to rely on a pathway for object vision, the "what" pathway and one for spatial vision, the "where" pathway [1]. A refinement of this concept emphasized the relevance of the "where" pathway for action control [2]. Almost all computational models of visual perception and attention follow this separation between "where" and "what". The general idea is, that the dorsal pathway first selects the location of an object and then the ventral pathway recognizes it by analyzing only a spatially defined part of the scene [3]. This decoupling of recognition and selection has the advantage of a facilitated recognition as compared to a fully parallel approach, since it is not practicable to apply several object models at the same time at several locations [4]. However, such a model of perception has its limitation if we search for a specific object. How could the "where" pathway know what is relevant?
The relevance of an object seems to be reflected by the activity of IT cells [5] [6]. Although the initial activation of IT neurons is largely stimulus driven and cells encoding target and non-target become activated, different populations compete for representation and typically the cells encoding the non-target are suppressed. Such competition is assumed to be biased by top-down feedback from working memory [5] [6]. A computational approach by Usher and Niebur [7] shows that

a parallel competition based on lateral interactions is sufficient to qualitatively replicate some of those findings, but they argue that the parallel stage is useless in case of a search for a conjunction and the decision has to be based on a serial scan of all objects.
It was suggested that the frontal eye field (FEF) could implement a saliency map by the convergence of information from different brain areas [8]. This raises the question how the FEF knows what is task relevant and where the object of interest is located. The FEF has connections to occipital, temporal and parietal areas, the thalamus, superior colliculus and prefrontal cortex [9]. The projections from V2 and V3 are weak, from V4 intermediate and heavy from TEO. Anterior IT cortex does not project directly to FEF. Information about the target features could be received from prefrontal areas and compared with features of intermediate complexity from V4 and TEO. This would require that the FEF or related areas perform a match detection in topological and topographic space. Alternatively, Desimone and Duncan [10] speculate "at some point in time, mechanisms for spatial selection may also be engaged to facilitate localization of the target for the eye movements". Some authors proposed feature specific top-down influences [11] [12] that could guide attention before the eye movement is planned. However, their implementation and exact function remained mysterious. Others suggested a top-down directed beam within the ventral pathway [13]. Only recently the influence of top-down feedback is beginning to be investigated more closely [14] [15] [16] [17] [18] [19]. In this paper I suggest that the visual areas process incoming stimuli first in a parallel bottom-up manner without a significant bottleneck and then acquire a more detailed knowledge about an object of interest by feedback. I show that such feedback within the ventral pathway can account for goal directed covert and overt search. Even for conjunction search a serial scan is not imperative.

2 Model

I model aspects of the areas V4, IT, FEF and PF and refer to the model by the prefix M (Fig. 1). M-IT, M-V4 and M-PF are subdivided into different dimensions (e.g., color and shape). My model consists of ascending populations, called (s) stimulus cells that can be primed by feedback connections and descending populations (t) target cells that project the dominant patterns back into the source areas.
The model prefrontal cortex serves for two major functions, memorizing a pattern in M-PFwm (working memory) cells and indicating a match of the incoming pattern with the memorized pattern in M-PF match cells. Thus, M-IT cells can only drive M-PFm cells when their pattern matches the prior knowledge from M-PFwm cells.
The neurons in the FEF can be categorized based on their responses to visual stimuli or to saccade execution into visual, visuomovement, fixation and movement cells [20]. I consider (v) visuomovement, (f) fixation and (m) movement cells in my model (Fig. 1). The M-FEFv neurons receive convergent afferents

from features in M-V4 at the same retinotopic location and add-up across all dimensions. M-FEFf cells generally inhibit M-FEFm cells. A threshold detection of the M-PF match cells is applied to determine if the target is in the search array. In this case the input into the M-FEFf cell is removed and thus the mapping from sensory to motor is facilitated. M-FEFv cells activate M-FEFm cells by surround inhibition. Since there is evidence that saccades are elicited when movement related activity in the FEF reaches a particular level [21], I assume a fixed threshold in M-FEFm cells to initiate a saccade. A spatially organized gain control input of M-V4 and M-IT stimulus cells originates from from M-FEFm cells.
M-PFwm cells modulate visual processing via feedback into M-ITs according to the current goal of the task. The resulting local increase of firing in M-ITs cells is directed further downwards by feedback form M-ITt cells to M-V4s cells. Thus, increased local activity in M-V4 enhances the visually responsive neurons in the frontal eye field, such that these cells reflect the task-relevance of a location.

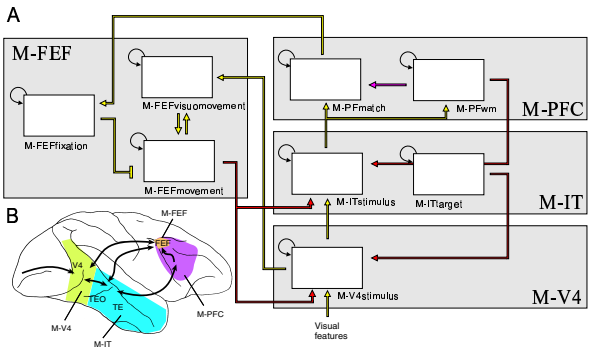


Fig. 1. (A) Sketch of the simulated areas. Each box represents a population of cells. The activation of those populations is a temporal dynamical process. Bottom-up (driving) connections are indicated by a bright arrow and top-down (modulating) connections are shown as a dark arrow. (B) Outline of the minimal set of interacting brain areas. Our model areas are restricted to elementary but typical processes and do not replicate all aspects of these areas.

3 Results

In order to demonstrate the possible role of feedback in the ventral pathway I simulated a memory guided search task [6] (Fig. 2A). If the same cued object reappears in the search array, the condition is called 'Target Present'. In the 'Target Absent' condition the cue stimulus is different from the stimuli in the choice array. Now a saccade has to be withheld.
The target was presented to the model and its features have been memorized in M-PFwm cells. Prior to the onset of the search array the active M-PFwm cells

increase the baseline activity of the M-IT cells selective for the target (Fig. 2B). When the search array appears, inputs are processed bottom-up without any strong bottleneck. Each cell initially encodes the presence of its preferred stimulus, but the target cell shows an early advantage due to top-down modulation from M-PFwm cells. Between 150 and 300 ms the cells encoding the non-target get suppressed although the input is still present, whereas the cells encoding the target remain active. A crucial condition is the target absent condition. Both non-targets decrease their activity, but less than in the distractor suppression case. A simple winner-take-all competition would not replicate the experimental data because due to noise in the system, a non-target would be selected in the target absent condition. My simulation results even match the temporal course of activity of IT cells in the different conditions of the experiment from Chelazzi et al. [6]. This constraint allows me to give reliable predictions of the processing in other areas.

The model predicts that the early advantage of IT cells encoding the target is sent to V4 cells, which have smaller RFs and creates an early target effect in V4 (see also [15]). Recent cell recordings confirmed this prediction: During the early phase until 150 ms after array onset, V4 cells show a slight target effect, which is stronger when two stimuli are located within a V4 receptive field [6].

Since FEFv neurons receive their main input from M-V4 an enhancement within the topographic/topological(feature) space is transferred into topographic space, such that a target selection is possible. This result explains how the visual cells of the FEF might discriminate over time the target from the distractor in conjunction visual search. The advantage in different dimensions adds up. The location of the target receives enhanced input from both dimensions. Locations encoding distractors sharing a single feature with the target receive enhanced input just from one dimension. The temporal course of activity of the FEFv and FEFm cells is similar to what has been found in experiments [8] [24]. FEFm cells quickly discriminate the target from the non-target.

The fronal eye field and areas within the dorsal pathway form a fronto-parietal network. These areas can use such a discrimination for overt and covert search. In overt search an eye movement is executed when the activity of the FEF movement cell reaches a threshold. Covert search is possible if activity, e.g. from the movement cells, reenters extrastriate visual cortex and enhances the input gain in V4 and IT in a spatially organized manner.

4 Discussion

This study demonstrates how findings in single cell recordings can be used to constrain models of perception. Each modeled area exhibits a temporal course of activity that has been observed by similar physiological experiments performed by various investigators. What are the major findings and predictions of this study for modeling object recognition and attention? First of all, the ventral pathway encodes an object of interest as well as its location. The model predicts that one role of feedback is to enhance the gain of cells encoding features of the

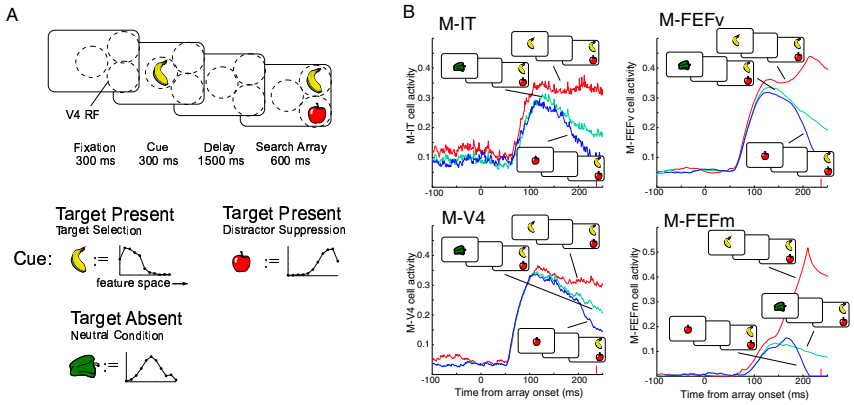


Fig. 2. (A) Simulation of the experiment of Chelazzi et al. [6]. The objects are represented by a noisy population input, here illustrated by a snapshot. RF's without an object just have noise as input. Each object is encoded within a separate RF, illustrated by the dashed circle, of M-V4 cells in two simulated dimensions (only one is shown). All M-V4 cells are within the RF of the M-IT cell population. The model has to indicate a successful search, by selecting the previously shown object as the target of an eye movement. (B) Activity within the model areas aligned to the onset of the search array in the different conditions.

object of interest. Such a mechanism would allow for a foreground-background discrimination throughout the ventral pathway down to V1.
Second, object recognition and attention recruit the same neural architecture. Recognition is related to the firing of detector cells and attention is typically implemented by control units. My model does not contain any control units. Competition and cooperation within the recognition network implements a dynamic filter that allows the brain to connect planning processes with the physical world. As a result, suppressive and facilitatory effects occur, commonly referred to as "attention".

Acknowledgements: This research was supported by DFG HA2630/2-1 and in part by the NSF (ERC-9402726).

References

1. Mishkin, M., Ungerleider, L.G., Macko, K.A.: Object vision and spatial vision: Two cortical pathways. Trends in Neurosci. **6** (1983) 414–417.
2. Milner, A.D., Goodale, M.A.: Visual pathways to perception and action. Progress in Brain Research **95** (1993) 317–337.
3. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Psychology **4** (1985) 219–227.
4. Ballard, D.H, Brown, C.M.: Principles of animate vision. In: Aloimonos, Y. (eds.): Active Perception. Lawrence Erlbaum Associates (1993) 245–282.

5. Chelazzi, L., Miller, E.K., Duncan, J., Desimone, R.: A neural basis for visual search in inferior temporal cortex. Nature **363** (1993) 345–347.
6. Chelazzi, L., Duncan, J., Miller, E.K., Desimone, R.: Responses of neurons in inferior temporal cortex during memory-guided visual search. J. Neurophysiol. **80** (1998) 2918–2940.
7. Usher, M., Niebur, E.: Modeling the temporal dynamics of IT neurons in visual search: A mechanism for top-down selective attention. J. Cog. Neurosci. **8** (1996) 311–327.
8. Thompson, K.G., Bichot, N.P., Schall, J.D.: From attention to action in frontal cortex. In: Braun, J., Koch, C., Davis, J.L. (eds.): Visual Attention and Cortical Circuits. MIT Press, Cambridge (2001), 137–157.
9. Schall, J.D., Morel, A., King, D.J., Bullier, J.: Topography of visual cortex connections with frontal eye field in macaque: Convergence and segregation of processing streams. J. Neurosci. **15** (1995) 4464–4487.
10. Desimone, R., Duncan, J.: Neural mechanisms of selective attention. Annu. Rev. Neurosci. **18** (1995) 193–222.
11. Treisman, A., Sato, S.: Conjunction search revisited. J. Exp. Psychol. Hum. Percept. Perform. **16** (1990) 459–478.
12. Wolfe, J.: Guided search 2.0 A revised model of visual search. Psychonomic Bulletin & Review **1** (1994) 202–238.
13. Tsotsos, J.K., Culhane, S.M., Wai, W., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning. Artificial Intelligence **78** (1995) 507–545.
14. Koechlin, E., Burnod, Y.: Dual population coding in the neocortex: A model of interaction between representation and attention in the visual cortex. J. Cog. Neurosci. **8** (1996) 353–370.
15. Hamker, F.H.: The role of feedback connections in task-driven visual search. In: Heinke, D., Humphreys, G.W., Olson, A. (eds.): Connectionist Models in Cognitive Neuroscience. Springer Verlag, London (1999), 252–261.
16. Hamker, F.H.: Distributed competition in directed attention. In: Baratoff, G., Neumann, H. (eds.): Dynamische Perzeption, Proceedings in Artificial Intelligence, Vol. 9. AKA, Akademische Verlagsgesellschaft, Berlin (2000) 39–44.
17. van der Velde, F., de Kamps, M.: From knowing what to knowing where: modeling object-based attention with feedback disinhibition of activation. J. Cogn. Neurosci. **13** (2001) 479–491.
18. Corchs, S., Deco, G.: Large-scale neural model for visual attention: integration of experimental single-cell and fMRI data. Cereb. Cortex **12** (2002) 339–48.
19. Roelfsema, P.R., Lamme, V.A., Spekreijse, H., Bosch, H.: Figure-ground segregation in a recurrent network architecture. J. Cogn. Neurosci. **14** (2002) 525–537.
20. Schall, J.D., Hanes, D.P., Thompson, K.G., King, D.J.: Saccade target selection in frontal eye field of macaque. I. Visual and premovement activation. J Neurosci **15** (1995) 6905–6918.
21. Hanes, D.P., Schall, J.D.: Neural control of voluntary movement initiation. Science **274** (1996) 427–430.
22. Hamker, F.H. Attention as a result of distributed competition. Soc. Neurosci. Abstr., Vol. 27, Program No 348.10, 2001.
23. Chelazzi, L., Miller, E.K., Duncan, J., Desimone, R.: Responses of neurons in macaque area V4 during memory-guided visual search. Cereb Cortex **11** (2001) 761–772.
24. Bichot, N.P., Rao, S.C., Schall, J.D.: Continuous processing in macaque frontal eye cortex during visual search. Neuropsychologia **39** (2001) 972-982.

Exogenous and Intention-Dependent Control of Attention Shifts in Dynamic Displays

Ingrid Scharlau and Ulrich Ansorge

Department of Psychology, Bielefeld University, P.O. Box 10 01 31, 33501 Bielefeld[1]
{ingrid.scharlau, ulrich.ansorge}@uni-bielefeld.de

Abstract. Two experiments investigated the control of attention shifts. Exogenous orienting [1], singleton capture [2], contingent orienting [3], and direct parameter specification [4] served as alternative hypotheses. Attentional allocation was assessed via its facilitating influence on perceived latency of stimuli. Facilitation was larger for intention-matching than for non-matching masked stimuli. This result tentatively supports the direct parameter specification account which predicts that masked visual information may directly specify open parameters of a response to the extent that they match intended features.

1 Introduction

Control of attention shifts in dynamic visual displays may be of several different types, such as *exogenous* or *bottom-up capture*, or *endogenous* or *volitional orienting* towards relevant stimuli matching the current intentions. According to the *attentional capture* account, sudden changes of peripheral stimulation elicit involuntary, stimulus-driven orienting towards the location of these changes [1, 5]. Recently, several alternative types of top-down control have been proposed. Folk, Remington, and Johnston [2] observed that onset cues did not capture attention if observers did not search for onset targets. They reasoned that attentional settings for specific feature classes controlled orienting in a top-down manner (*contingent capture*). They further observed limitations with respect to the features that can be specified in attentional sets: Control settings can be directed to either dynamic features, such as abrupt onset and motion, or static features such as specific colours. However, if observers are set for abrupt onset targets, other dynamic features, such as motion targets, will also capture attention.

An alternative top-down approach is the *direct parameter specification* model (DPS) [4, 6]. It likewise proposes that stimuli may control attention only to the extent that they match intended features. However, types of features apt for direct processing are not restricted. Additionally, DPS explicitly allows for not consciously perceived information to exert control over responses. The DPS concept was originally developed while studying sensorimotor effects of masked stimuli [7]. It assumes that, as far as an

[1] The research reported in this paper was supported by the Deutsche Forschungsgemeinschaft (DFG), Grant NE-366/5-2 to Odmar Neumann.

action plan is available, response parameters can be specified by direct processing pathways from stimulus to response that bypass a conscious representation. Masked visual information indeed has been shown to lead to the activation of a corresponding response [8, 9]. Control of attention shifts might be another case of DPS with the parameters specified being the amplitude and the direction of an attention shift.

A third type of intention-dependent control has been proposed by Bacon and Egeth [1]. They reported that if observers search for a feature singleton (a single deviating feature), other singletons may interfere with visual search even if they do not contain the relevant features. They thus distinguish between two top-down controlled search strategies, singleton search (set for any singleton present) and feature search (set for any stimulus that has a certain feature).

The present study explored the contributions of these types of orienting to perceived latency of visual stimuli. Distribution of attention over the visual field was assessed by means of *perceptual latency priming* (PLP). In PLP, the latency of a stimulus is decreased by a masked prime that precedes it. PLP results from an attention shift towards the prime's location which facilitates processing of the trailing target. Earlier studies of PLP revealed evidence for the contribution of exogenous orienting. For example, PLP has been found to be independent of similarity between prime and target [10]. By contrast, in a recent study, larger effects for intention-matching than for non-matching primes were found [6]: The primes were either similar to the targets or similar to irrelevant distractor stimuli. Target-like, but not distractor-like, primes facilitated perceptual latencies of targets trailing at their positions supporting the DPS account. However, the results of this study were also in line with an explanation by singleton capture: Observers searched for singleton targets, and the target-like prime may thus have captured attention due to a singleton-detection strategy. This was not possible for the distractor-like prime since it was always preceded by at least one similar distractor.

2 Method

Throughout the experiments, PLP was assessed by temporal order judgments (TOJ). Participants judged the temporal order of two targets in a small set of distractors. One of the targets could be primed by a smaller stimulus (a prime). Size and temporal sequence of prime and target stimuli met the conditions of metacontrast masking [11]. From the psychometric distributions of order judgments, Points of Subjective Simultaneity (PSS) were computed by logit analysis [12]. PLP was measured by differences between primed and unprimed PSS values. Discrimination performance was measured by mean slope of the inner quartile of the psychometric distributions (Difference Limen, DL). If necessary, degrees of freedom were corrected by the Greenhouse-Geisser coefficient, and adjusted alpha values are given [13].

If PLP is due to *exogenous orienting* towards the location of the prime, it will be independent of whether the primes resemble target features. On the other hand, it will be influenced by prime validity, that is, the extent to which a prime predicts the location of a subsequent target [13]. If PLP is due to *singleton capture*, it will arise exclusively if the participants have the opportunity to search for singletons, and the prime

is a feature singleton. If *DPS* is responsible for PLP, intention-matching, though not non-matching, primes will attract attention. By contrast, if the control of attention in TOJ tasks is due to *contingent orienting*, there will be no difference between matching and non-matching primes which differ within a static feature since participants are set to search for a dynamic feature (onset).

Two main experimental factors, *prime match* and *prime validity*, serve to decide between these alternative accounts. Prime match was manipulated by presenting primes that resembled either the targets (matching prime) or the distractors (non-matching prime). Prime validity was manipulated by the number of primes presented. In the valid case, a single prime was presented at the location of one target, and in the neutral condition, two primes appeared simultaneously, one at a target location and the other one at an otherwise blank location. According to the exogenous-orienting account, a single valid prime will have a larger effect than a prime that is presented simultaneously with a competing stimulus. This manipulation of validity also allowed to control for the influence of singleton capture since only the valid prime was a feature singleton. The manipulation of prime match served to differentiate between the DPS and the other accounts since only the former predicts an exclusive influence of matching primes in the PLP paradigm.

We controlled the observers' task strategy by presenting two different targets or two similar targets in different blocks. In the latter case, observers adopt a feature-search strategy whereas the former case allows a singleton-detection strategy since each of the targets is a singleton. According to the exogenous-orienting and singleton-detection account, though validity effects may be absent in feature-detection mode, primes will have an influence on PLP in singleton-search mode.

3 Experiment 1

Participants judged the temporal order of two targets while disregarding additional visual distractors. In one of the two sessions, they performed the task in singleton-search mode, in the other one, feature-search mode was forced. 16 voluntary participants with a mean age of 25 years took part in the experiment. All had normal or corrected-to-normal vision.

Stimuli were red, yellow, and blue rings on dark grey background. In each trial, four non-offset visible rings were presented equidistant to fixation, two distractors defined by colour, and two targets also defined by specific colours. Intervals between the targets were 192, 128, and 64 ms. Prime stimuli were smaller rings. The prime (presented for 32 ms) preceded one of the targets by 64 ms. In the *matching* condition, it had the same colour as the masking target, whereas in the *non-matching* condition, it had the distractor colour. In the *valid* condition, one prime was presented at a location subsequently occupied by a target. In the *neutral* condition, a second prime was simultaneously presented at a further unoccupied location. As a baseline, unprimed trials were included. In *singleton-search mode*, targets had different colours so that each target was a feature singleton. Observers indicated which target colour had been the first one. In *feature-search mode*, targets had the same colour. After presentation of the trial, one of them was marked and observers had to indicate if it had appeared

first or second. Apart from the unprimed baseline condition, there were 8 conditions (2 tasks × 2 prime match conditions × 2 prime validity conditions).

Experiment 1 revealed a priming effect on PSS: The prime facilitated perception of the primed target by an average of 20 ms (see Fig. 1). However, PLP did not differ due to the experimental factors (main effects and interactions: $F < 1$). Separate t-Tests of PLP values for each condition revealed that with one exception (feature search / non-match / neutral condition), all latencies differed significantly from zero (all $p < .00625$). Discrimination performance was lower in feature search than in singleton search ($F[1, 16] = 6.25$; $p < .05$). It was slightly lower with a non-matching than with a matching prime ($F[1, 16] = 3.98$; $p = .0634$) and with neutral compared with valid primes ($F[1, 16] = 3.76$; $p = .0703$).

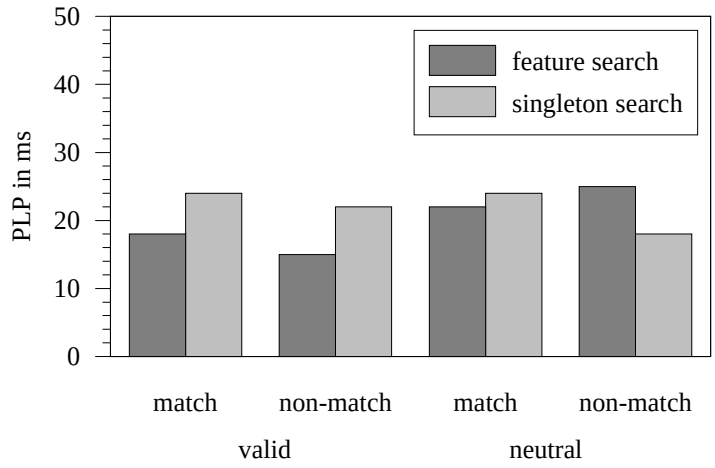


Fig. 1. PLP values in Experiment 1. The combinations of the main experimental factors (prime match and prime validity) are given on the abscissa, the two tasks as separate columns

The DL results indicate that the visible distractor prime may have interfered with TOJ. The valid and the neutral condition were not strictly comparable since the valid prime was masked by the trailing target whereas the neutral prime was not and may have led to a confusion of prime and target. This was controlled for in Experiment 2.

4 Experiment 2

Experiment 2 replicated Experiment 1 with the single exception that both primes in the neutral condition were masked, the second one by a distractor trailing at its location. There were 16 voluntary participants with a mean age of 27.6 years. All had normal or corrected-to-normal vision.

Again, PLP was found. On average, it was 16 ms (see Fig. 2). Search strategy had no effect on PLP ($F[1, 16] = 2.46$; $p = .1367$), as well as prime validity ($F < 1$). Matching primes entailed larger PLP effects than non-matching primes ($F[1, 16] =$

13.78; $p < .01$), a finding which was qualified by a task × match interaction ($F[1, 16] = 5.5$; $p < .05$). The differential effects of matching and non-matching primes were larger in the singleton-search task than in the feature-search task. Separate t-Test of PLP values for each condition revealed that three PLP values differed significantly from zero: the feature search / matching / valid condition (PLP: 16 ms), and the singleton search / matching / valid (25 ms) as well as singleton search / matching / neutral condition (30 ms; all $p < .00625$). DL was slightly larger in feature search than in singleton search ($F[1, 16] = 8.48$; $p < .05$). No further influences on DL were found. In sum, Experiment 2 revealed an advantage of matching primes in the control of attention shifts.

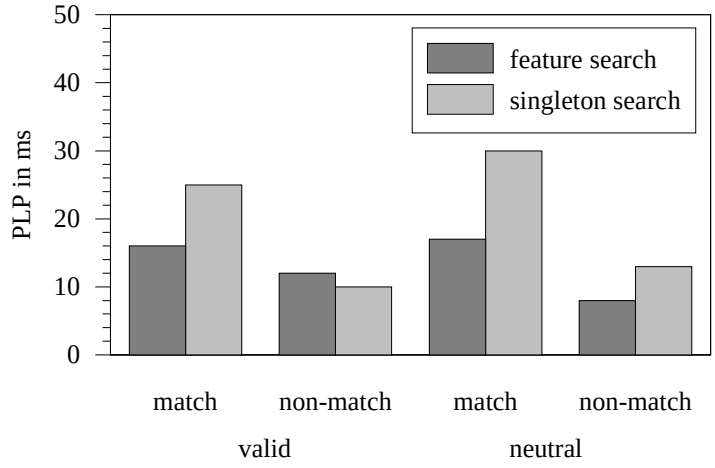


Fig. 2. PLP values in Experiment 2. The combinations of the main experimental factors (prime match and prime validity) are given on the abscissa, the two tasks as separate columns

5 General Discussion

The absence of a prime validity effect on PLP in the experiments is not in line with an exogenous-orienting account: According to this explanation, presenting the prime simultaneously with a second prime impairs its effect on orienting of attention since the primes compete for capture. It also disfavours a singleton-capture account. If subjects searched for feature singletons in the TOJ task and the prime captured attention due to its being a singleton, again no effects in the neutral condition would have been expected.

Some support for top-down control of attention is revealed by Experiment 2: Differential effects of matching and non-matching primes were found in Experiment 2, as predicted by the DPS account. This effect is not predicted by the contingent orienting account. The control of attention shifts thus seems to be possible in a mode of DPS. However, the influence of non-matching primes on PLP found in Experiment 1 indi-

cates an additional effect of an irrelevant stimulus. This capture effect may be involuntary, or it may be due to other top-down processes, such as whether target and distractor features are linearly separable [15, 16]. Comparison with earlier studies [10] reveals that the priming effect in the present study is rather small. With priming intervals of 64 ms, PLP typically amounts to about half of this interval. With an average of 28 and 24 ms, the priming effect of matching primes was substantially smaller even in the singleton-search sessions; in the feature-search sessions, it was further reduced to 17 and 20 ms. This may be due to an increment in task difficulty that could have left less space for differential effects of intended and non-intended signals to show up.

References

1. Jonides, J.: Voluntary versus automatic control over the mind's eye's movement. In: Long, J.B., Baddeley, A.D. (eds.): Attention and Performance IX. Erlbaum, Hillsdale, NJ (1981) 187-203
2. Bacon, W.F., Egeth, H.E.: Overriding stimulus-driven attentional capture. Percept Psychophys 55 (1994) 485-496
3. Folk, C.L., Remington, R.W., Johnston, J.C.: Involuntary covert orienting is contingent on attentional control settings. J Exp Psychol Human, 18 (1992) 1030-1044
4. Neumann, O.: Direct parameter specification and the concept of perception. Psychol Res-Psych Fo, 52 (1990) 207-215
5. Yantis, S., Jonides, J.: Abrupt visual onsets and selective attention: Evidence from visual search. J Exp Psychol Human, 10 (1984) 601-621
6. Scharlau, I., Ansorge, U.: Direct parameter specification of an attention shift: Evidence from perceptual latency priming. Manuscript submitted for publication (2002)
7. Klotz, W., Neumann, O.: Motor activation without conscious discrimination in metacontrast masking. J Exp Psychol Human, 25 (1999) 976-992
8. Eimer, M., Schlaghecken, F.: Effects of masked stimuli on motor activation: Behavioral and electrophysiological evidence. J Exp Psychol Human, 24 (1998) 1737-1747
9. Leuthold, H., Kopp, B.: Mechanisms of priming by masked stimuli: Inferences from event-related brain potentials. Psychol Sci, 9 (1998) 263-269
10. Scharlau, I.: Leading, but not trailing primes influence temporal order perception: Further evidence for an attentional account of perceptual latency priming. Percept Psychophys (in press)
11. Breitmeyer, B.G.: Visual masking: An integrative approach. Oxford University Press, Oxford UK (1984)
12. Finney, D.J.: Probit analysis. 3rd edn. University Press, Cambridge MA (1971)
13. Hays, W.L.: Statistics 4th edn. Holt, Rinehart, and Winston, Orlando, FL (1988)
14. Posner, M. I.: Orienting of attention. Q J Exp Psychol, 32 (1980) 3-25
15. Bauer, B., Jolicoeur, P., Cowan, W.B.: Distractor heterogenity versus linear separability in color visual search. Perception 25 (1996) 1281-1293
16. Hodsoll, J., Humphreys, G.W.: Driving attention with the top down: The relative contribution of target templates to the linear separability effect in the size dimension. Percept Psychophys, 63 (2001) 918-926

Hemispheric Asymmetries for Global/Local Processing in Varied Mapping Tasks

Gregor Volberg, Ronald Hübner

Universität Konstanz, Fachbereich Psychologie, Fach D29, D-78457 Konstanz, Germany
gregor.volberg@uni-konstanz.de

Abstract. There is some neuropsychological evidence for a differential capacity of the cerebral hemispheres to process local and global levels of compound visual stimuli. Corresponding visual field (VF) effects in response time studies, though, are mainly obtained with stimuli that induce response conflicts with respect to the levels. Here we investigate why response conflicts are favorable to VF–effects. Two experiments with hierarchical letters are reported, in which the difficulty of response selection was varied for conflicting and non-conflicting stimuli. For the difficult situation, VF–effects were also obtained for non-conflicting stimuli. The results are interpreted in the way that in both cases the letter identity and the corresponding stimulus level had to be integrated.

The human brain is subdivided into two homologous areas, the left and right cerebral hemisphere, which perform some cognitive functions with different efficiency. One example is the differential hemispheric capacity to process large-scaled (i.e., global) and small-scaled (i.e., local) aspects of compound visual objects. This asymmetry was often reported in studies with brain-damaged patients, where right- and left-hemispheric lesions were accompanied by impairments for the processing of global and local stimulus aspects, respectively [1].

Corresponding hemispheric differences in response time studies, though, are only obtained if a number of favorable conditions are met [2]. One such condition that turned out to be particularly important is a response conflict between the information on the global and that on the local level of the stimulus [for a meta-analysis see 3]. For instance, Hübner and Malinowski [4] presented compound stimuli to the left visual field/right hemisphere (LVF/RH) or right visual field/left hemisphere (RVF/LH), and let their subjects name the form on the global or local level. In all three experiments they conducted, Hübner and Malinowski only found an interaction between visual field and target level for those stimuli where the global and local information was mapped to different responses.

To explain this effect, Hübner and Malinowski suggested that response selection for non-conflicting and for conflicting stimuli is performed in qualitatively different modes, respectively. They argued that for the former type, fast and automatic responses can be released before the hierarchical structure of the stimulus is represented. This could be accomplished with equal efficiency in the LH and RH. Contrarily, a more controlled mode of response selection is required for conflicting stimuli. Here, subjects must integrate the global and local forms with the corresponding stimulus level in order to select a correct answer. Hübner and Malinowski suggested that

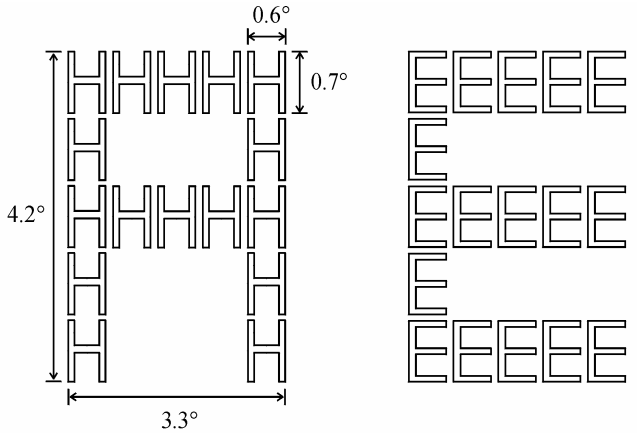


Fig. 1. Two examples of compound letters, where the global shape is composed by local elements in a 5 by 5 grid. Four letters (A, S, H, and E) were used and combined to 16 hierarchical stimuli. The size of the local and global letters is given in degrees of visual angle

this binding process is performed with different efficiency in the left and right hemisphere.

Unfortunately, there is as yet no clear evidence for this hypothesis. A major drawback of the reported study was that hemispheric asymmetries were exclusively obtained with conflicting stimuli. Consequently, it can not be ruled out that conflicting responses are necessary to produce these effects [3]. If, however, the mode of response selection is crucial to hemispheric asymmetries, then it should be possible to induce them by means other than response conflicts. This prediction was tested in the present study. Two experiments were conducted, where response conflicts and the mode of response selection were varied independently. To achieve this, the assignment of stimuli to response keys was held variable. The underlying rationale was that a varied mapping procedure would hinder subjects from giving automatic responses, because a more thorough evaluation of the stimulus must be performed to select the correct answer [5]. Under this constraint, we expected that hemispheric asymmetries would be obtained with conflicting as well as non-conflicting stimuli.

Experiment I

Eight right-handed volunteers (4 female, 4 male, aged 22-30 years) participated in this experiment. They performed 16 blocks of 32 trials within one experimental session. The trials started with a central 300 *ms* presentation of a cue that indicated the target level for the following stimulus. After a cue-stimulus-interval of 300 *ms*, the subjects were presented with hierarchical letters [6], which appeared in the LVF or RVF for 93 *ms* (for a description of the used stimuli see Figure 1). Between the response and the following trial, there was an interval of 1000 *ms*. The task was to categorize the letter at the cued level of the hierarchical stimulus by pressing the left or right button of a response device.

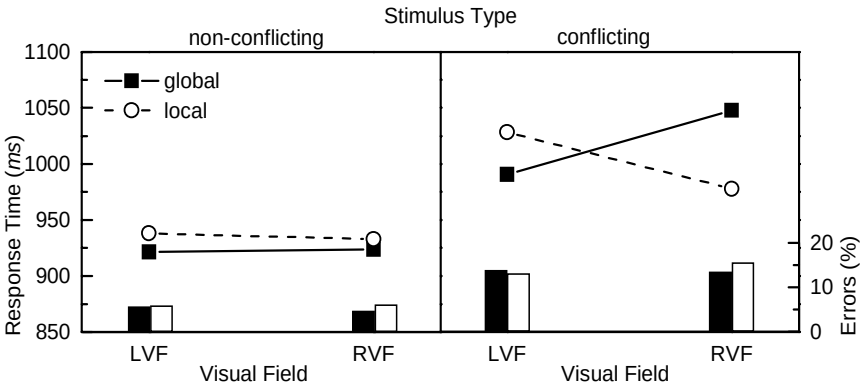


Fig. 2. Interaction between target level, visual field, and stimulus type as revealed in the first experiment

As in the Hübner and Malinowski study, half of the presented stimuli were conflicting, whereas the other half was non-conflicting. The four letters used were grouped to two response categories, which were mapped to the left and right response key, respectively. However, only one letter within each category was consistently mapped to a fixed response, whereas the mapping of the other letters was changed after each block. For example, the letters *A* and *H* could have a fixed mapping, whereas the mapping of the letters *S* and *E* was variable. In this case, the mapping of letters to the respective left/right response in succeeding blocks was AS/HE, AE/HS, AS/HE, AE/HS and so forth.

Because stimulus-response mappings were changed frequently, it was unlikely that automatic responses would develop [7]. We thus expected that subjects would apply a more controlled mode of response selection, where form and level of the hierarchical stimulus are integrated. Accordingly, the hypothesis was that hemispheric asymmetries would be obtained with conflicting as well as non-conflicting stimuli. The factors in the first experiment were target level (global, local), visual field (LVF, RVF), stimulus type (conflicting, non-conflicting), and target mapping (fixed, variable), which were all randomized.

Results & Discussion

Error rates and latencies of correct responses were entered into an analysis of variance (ANOVA) with repeated measures on all factors. The focus in this as well as in the second experiment was on visual field (VF)-effects, that is, on the greater capacity of the LH and RH to process local and global stimulus aspects, respectively. In parallel to Hübner and Malinowski, we will express VF-effects for local elements by subtracting response latencies to RVF-stimuli from those to LVF-stimuli, and analogously VF-effects for global forms are given by subtracting response latencies to LVF-stimuli from those to RVF-stimuli. For both levels, thus, positive values indicate VF-effects in the expected direction.

Generally, responses were faster for non-conflicting compared to conflicting stimuli [929 *ms* vs. 1011 *ms*, $F(1,7) = 15.23$, $p < .01$], and for targets with fixed compared to variable mapping [928 *ms* vs. 1012 *ms*, $F(1,7) = 21.14$, $p < .01$]. Similar effects were also revealed with the error rates. Reliable VF-effects, though, were only obtained for response latencies. First, there was a two-way interaction between target level and visual field [$F(1,7) = 11.69$, $p < .05$]. However, this was qualified by a three-way interaction between target level, visual field, and stimulus type [$F(1,7) = 9.02$, $p < .05$]. The corresponding results are depicted in Figure 2. As one can see, the expected interaction between target level and VF held for conflicting stimuli [$F(1,7) = 19.24$, $p < .01$], but not for non-conflicting stimuli [$F(1,7) = 0.12$, $p = .74$]. The interaction for the former type was due to large, though non-significant, global and local VF-effects (57 *ms* and 50 *ms*, respectively). The corresponding (non-significant) effects for non-conflicting stimuli were 2 *ms* and 5 *ms*, respectively.

The above results did obviously not meet our hypothesis. One possible explanation for this flaw is that the subjects could establish automatic responses despite the varied mapping. This might have been favored by the fact that only half of the presented letters were indeed mapped to variable responses. Moreover, the response mapping was only changed after each block. To account for this possible shortcoming, a second experiment was conducted, where all four letters were mapped to variable responses. As well, the mapping was changed within the blocks.

Experiment 2

16 right-handed volunteers (12 female, 4 male, aged 19-27 years) took part in the second experiment. The procedure and the stimuli were basically the same as in experiment one. The main difference to the first experiment was the response mapping. Here, the four letters were grouped to two response categories (‘A, S’, ‘H, E’), which were consistently mapped to the left and right response key, respectively. However, the mapping rules were reversed for global and local targets. For example, the subjects had to press the left button for a global *A* or *S*, but the right button if *A* or *S* appeared at the local level. Accordingly, a global *H* or *E* required a right button press, whereas a local *H* or *E* were mapped to the left response key. The same letter was thus always mapped to two different responses. As a consequence, the subjects could not give a proper answer before the hierarchical structure of the stimulus was represented. This applied to non-conflicting stimuli as well as to conflicting stimuli. An exception from that were those stimuli with the same letters on both levels, e.g., a global *H* with local *H*s. Notice that this type of stimulus was conflicting, because the global and the local *H* were assigned to different responses. To illustrate the difference between conflicting stimuli with different letters (conflicting/d) and those with the same letters (conflicting/s) on the global and local level, consider a trial where the task was to categorize the local letter of the described example stimulus. It is clear that this local element could only be *H*, because there was no alternative letter in the compound stimulus. In contrast to conflicting/d stimuli, though, the response to the local letter could here be selected from an early, incomplete stimulus representation.

Table 1. Response latencies and visual field-effects (*ms*) to global and local targets in the second experiment. The last row shows the interaction between target level and visual field (VF, see results section for details)

Target and VF		Stimulus Type		
		non-conflicting	conflicting/d	conflicting/s
Global				
	LVF	756	757	720
	RVF	793	780	742
Local				
	LVF	926	915	908
	RVF	925	898	903
VF-effects				
	Global	37[c]	23	22[a]
	Local	1	17	5
	Global + Local	38[b]	40[a]	27

Note. [a] $p < .10$, [b] $p < .05$, [c] $p < .001$

Two hypotheses could be tested with the present experiment. The first is that response conflicts are necessary to induce hemispheric asymmetries in global/local processing. If so, then one should obtain respective VF-effects only with conflicting stimuli (conflicting/d and conflicting/s). The second hypothesis is that the hemispheres differ in their capacity to integrate the stimulus level and form. If this was true, then respective VF-effect should only show up with stimuli where such integration needs to be performed (non-conflicting and conflicting/d). The factors in the second experiment were target level (global, local), visual field (LVF, RVF), and stimulus type (non-conflicting, conflicting/d, conflicting/s), which were all randomized.

Results & Discussion

Latencies of correct responses and error rates were subjected to an ANOVA with repeated measures on all factors. As in the first experiment, reliable VF-effects were only obtained with response latencies. The corresponding results are depicted in Table 1. One can see that the global VF-effects were reliable for non-conflicting stimuli [37 *ms*, $F(1,15) = 17.30$, $p < .001$] and for conflicting/s stimuli [22 *ms*, marginally significant: $F(1,15) = 3.53$, $p = .08$]. The global VF-effect to conflicting/d stimuli was considerably high, but not significant [23 *ms*, $F(1,15) = 2.68$, $p = .12$]. As well, none of the local VF-effects was significant.

The most important results with respect to the hypotheses were interactions between target level and VF. When the data was collapsed over all stimuli, this interaction was reliable [$F(1,15) = 5.63$, $p < .05$]. However, planned comparisons revealed that this would not hold for all stimulus types. Thus, the results are given separately for non-conflicting stimuli, conflicting/d stimuli and conflicting/s stimuli (see last row of Table 1). Here, the size of the interaction is expressed as the sum of global and local VF-effects. The value is higher the larger the expected hemispheric differences

are. Conversely, the value is lower if the expected hemispheric differences are small, or if one or both of the VF-effects point in the unexpected direction.

In line with the results from former experiments, a reliable interaction between target level and VF was found for conflicting stimuli. This, however, only held for conflicting/d stimuli [marginally significant: $F(1,15) = 3.63$, $p = .08$], but not for conflicting/s stimuli [$F(1,15) = 1.43$, $p = .25$]. Mind that the interaction effect for the former type was considerably larger than that for the latter type (40 *ms* vs. 27 *ms)*. As a second major result, the expected hemispheric differences were also reliable with nonconflicting stimuli [38 *ms*, $F(1,15) = 6.79$, $p < .05$]. Both results are clearly in odd to the hypothesis that response conflicts are a necessary condition for hemispheric asymmetries in global/local processing. Contrarily, the results support the notion that both hemisphere differ in their capacity to integrate the form and the level of compound visual stimuli.

Conclusions

Both experiments showed again that hemispheric asymmetries for the processing of global and local stimulus aspects can be obtained if the subjects respond to conflicting stimuli. However, the data also suggest that response conflicts are not the only way to induce VF-effects. Rather, hemispheric asymmetries occurred also under other conditions that require a thorough stimulus evaluation, i.e. conditions where the stimulus level and form had to be integrated. Thus, the present data support Hübner and Malinowski's feature-integration account of hemispheric asymmetries in global/local processing.

References

1. Robertson, L.C., Lamb, M.R.: Neuropsychological contributions to theories of part/whole organization. Cognit. Psychol. 23 (1991) 299-330
2. Yovel, G., Levy, J., Yovel, I.: Hemispheric asymmetries for global and local visual perception: effects of stimulus and task factors. J. Exp. Psychol. Hum. Percept. Perform. 27 (2001) 1369-85
3. Van Kleeck, M.H.: Hemispheric differences in global versus local processing of hierarchical visual stimuli by normal subjects: new data and a meta- analysis of previous studies. Neuropsychol. 27 (1989) 1165-78
4. Hübner, R., Malinowski, P.: The effect of response competition on functional hemispheric asymmetries for global/local processing. Percept. Psychophys. (in press)
5. Shedden, J.M., Reid, G.S.: A variable mapping task produces symmetrical interference between global information and local information. Percept. Psychophys. 63 (2001) 241-52
6. Navon, D.: Forest before the trees: The precedence of global features in visual perception. Cognit. Psychol. 9 (1977) 353-393
7. Shiffrin, R.M., Schneider, W.: Automatic and controlled processing revisited. Psychol. Rev. 91 (1984) 269-276

Human movement analysis

A Multimodal Person Tracking System Based on a Variant of the Condensation Algorithm

Harald Breit and Gerhard Rigoll

Institute for Human-Machine-Communication
Technical University of Munich, Germany
http://www.mmk.e-technik.tu-muenchen.de/

Abstract. We present a system for tracking persons which is suited for environments with a constant as well as a moving background. It is based on a variant of the condensation algorithm and is capable of combining the outputs of several measurements, so that it can be considered as a multimodal tracking system. The measurement modes which are currently implemented are a pseudo 2-dimensional hidden Markov model (P2DHMM), a color based skin finder, and a motion detector. The purpose of the combination of several modes is to make the tracking system more robust in critical situations by combining the individual strengthes of different modes. The architecture of this tracking system is described and some exemplary results are depicted.

1 Introduction

The tracking of moving objects in video sequences is a major problem in the area of visual surveillance and vision-based man-machine-interfaces. We have proposed approaches where the main goal was the possibility to track persons in front of moving backgrounds. For this we used a combination of a pseudo 2-dimensional hidden Markov model (P2DHMM) and a Kalman filter (see e. g. [1, 5]). This combination delivered good results, and so the question arose how this approach could further be improved with regard to robustness and the possibility of handling occlusion effects.

Because it seems that each method for locating a desired object has its specific advantages and disadvantages, one could try to combine the advantages of different measurement methods and at the same time to overcome their special disadvantages. This leads to the idea of so-called *multimodal tracking methods*, where several modes are exploited in order to increase the robustness of a tracking algorithm under real-world conditions.

The two basic problems when developing a multimodal tracking system are firstly to select and implement the different modes and secondly to successfully combine this modes. The approaches that we investigated here are a combination of a P2DHMM with a skin finder or a motion detector for person tracking. The motivation for this choice was to sustain our proven P2DHMM system as one of the modes in our new multimodal system. As second mode a color based skin finder has been considered to be a good complementary information source, since

skin and face information is not explicitly considered in the P2DHMM (which operates on gray level images) and especially since this second mode would be suitable to recover the tracking process in case of occlusions of the lower body. As a third mode a motion detector has been used, which robustly works on image sequences with a constant background. The tracking modes are merged in a probabilistic way using the condensation algorithm [2]. The condensation algorithm has found to be especially interesting for multimodal fusion since it offers flexible methods for a stochastical combination of the conditional measurement probabilities which are generated by the different tracking modes.

2 Principles of the condensation algorithm

The purpose of this algorithm is to describe the temporal propagation of conditional densities, which can be decomposed into three temporal consecutive steps, namely a deterministic drift, a stochastic diffusion and a reactive effect of a measurement. This is also done e. g. by a Kalman filter, but the condensation algorithm has the advantage that it is simpler from a mathematical point of view and therefore allows an uncomplicated combination of several measurement modes, as will be shown later.

In the following text we denote the state of the modeled object at the discrete time k as $\mathbf{x}_k = \mathbf{x}(t_k)$ and its history as $\mathbf{X}_k = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$. In an analogous manner a set of image features is gathered in a measurement or observation vector $\mathbf{z}_k$ with the history $\mathbf{Z}_k = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\}$. Using these symbols and Bayes' rule the tracking problem can be formulated in terms of conditional probabilities:

$$p(\mathbf{x}_k|\mathbf{Z}_k) \propto p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{Z}_{k-1}) \tag{1}$$

The condensation algorithm uses a set of samples of the state vector to approximate its conditional probability density function $p(\mathbf{x}_k|\mathbf{Z}_k)$. This sample set consists of N samples $\mathbf{s}_k^{(n)}$, each weighted with the probability $\pi_k^{(n)}$ which is obtained from the measurement $p(\mathbf{z}_k|\mathbf{x}_k = \mathbf{s}_k^{(n)})$. Now the conditional state density can be represented by the weighted sample set $(\mathbf{s}_k^{(n)}, \pi_k^{(n)}, n = 1 \ldots N)$.

For a description of how this sample set can be obtained recursively from the previous sample set and for further details see e. g. [2].

3 Computation of the conditional probabilities

The conditional probabilities $\pi_k^{(n)}$ have to be acquired by a measurement within the current image. Our approach is currently able to utilize three methods for acquiring this measurement data, namely a P2DHMM, a skin finder, and a motion detector.

The problem is now to evaluate a measurement vector $\mathbf{z}_k$ which results from one of the measurement modes (delivering e. g. a bounding box) in such a way that we can compute the conditional probability of this measurement under the condition of a given sample, expressed as $p(\mathbf{z}_k|\mathbf{x}_k = \mathbf{s}_k^{(n)})$. The relation between

$\mathbf{z}_k$ and $\mathbf{x}_k$ is expressed by the measurement equation $\mathbf{z}_k = \mathbf{H}\cdot\mathbf{x}_k + \mathbf{v}_k$, where $\mathbf{H}$ is the measurement matrix and $\mathbf{v}_k$ is the measurement noise. If $\mathbf{v}_k$ is white noise, it is a reasonable assumption that the variable $\mathbf{z}_k$ is a stochastic process that can be characterized by a Gaussian distribution where $\mathbf{H}\mathbf{x}_k$ can be considered as mean value of the process. In this case the above mentioned Gaussian distribution can be interpreted as the probability of the measurement vector $\mathbf{z}_k$ under the assumption that the sample $\mathbf{s}_k^{(n)}$ is the correct state vector, resulting in

$$p(\mathbf{z}_k|\mathbf{x}_k = \mathbf{s}_k^{(n)}) \propto \exp(-\frac{1}{2}(\mathbf{z}_k - \mathbf{H}\mathbf{x}_k)^T \mathbf{C}(\mathbf{z}_k - \mathbf{H}\mathbf{x}_k)). \tag{2}$$

In this function $\mathbf{C}$ denotes the covariance matrix which has to be chosen appropriately. The resulting probabilistic values are subsequently normalized so they will sum up to 1.

The state vector $\mathbf{x}$ (and each sample vector $\mathbf{s}$) consists of the components $\mathbf{x} = [x_c, y_c, v_x, v_y, w, h]^T$, where x_c and y_c describe the center of a bounding box with the width w, the height h and the velocity components v_x and v_y.

The functionality of this approach can be confirmed easily by the following assumptions: If the current measurement vector $\mathbf{z}_k$ is almost identical to $\mathbf{H}\mathbf{x}_k$, then measurement and sample must be located very closely together (i. e. $\mathbf{z}_k$ confirms $\mathbf{x}_k$ very well) and thus (2) will yield a very high probability for this sample. It is therefore a suitable equation for the probabilistic interpretation of the output $\mathbf{z}_k$ of our various modes.

3.1 P2DHMM

The abbreviation P2DHMM stands for pseudo 2-dimensional hidden Markov model. We will describe this method only very briefly here; for further details see e. g. [4, 1, 3]. The model which we used consists of 20 states which are arranged in 4 superstates (modeling columns) with each of them containing 5 normal states. The model has been trained to several hundred images that each show just one person surrounded by some arbitrary complex background. After this training has been accomplished, an image containing a person can be presented to the P2DHMM, and by means of the Viterbi algorithm the most probable state sequence and assignment of states to image areas can be calculated. In this way one obtains a segmentation of the image into person and background blocks. From this segmentation a bounding box (the smallest rectangle with horizontal and vertical edges that contains all pixels classified as person) and its center can be extracted.

Furthermore, the velocity of this bounding box can be calculated as the difference of the position of the center of the bounding box in the current frame and its position in the previous frame. Because this value can be very volatile, we smooth it by calculating a weighted mean value of the current velocity (70 %) and the previous velocity (30 %). Thus the result of the measurement of the P2DHMM will be a measurement vector of the form $\mathbf{z}_{\mathrm{P2D}} = [x_c, y_c, v_x, v_y, w, h]^T$, and the appropriate measurement matrix is a unity matrix.

3.2 Skin finder

As a second method for acquiring measurement data we use a simple implementation of a skin finder. The intention here was not to optimize this skin finder, but to demonstrate how a second measurement can be integrated into our condensation based tracking approach. As will be shown later, this measurement can have a strong positive influence on the tracking results, even if it is not always very accurate.

The skin finder is based on an approach using color histograms and conditional probabilities as it is described e. g. in [6]. The result of this measurement will be a two dimensional vector which describes the center of gravity of the skin colored pixels and has the form $\mathbf{z}_{\text{skin}} = [x_{\text{cog,skin}}, y_{\text{cog,skin}}]^T$. Because this point is expected to indicate the position of the face of a person, it will be positioned somewhat higher than the center of the bounding box by an amount which can be estimated to be approximately 30% of the height of the bounding box. Therefore, for the measurement matrix of the skin finder we use

$$\mathbf{H}_{\text{skin}} = \begin{bmatrix} 1\,0\,0\,0\,0 & 0 \\ 0\,1\,0\,0\,0 & -0.3 \end{bmatrix}. \tag{3}$$

3.3 Motion detector

As a third method for acquiring measurement data we use a motion detector. Again here the intention was to demonstrate how another measurement can be integrated into our condensation based tracking approach and thus improve the tracking results. The motion detector bases on a calculation of differences d between pixels $\mathbf{i}(x, y)$ in the current image and corresponding pixels in a reference image according to

$$d_k(x, y) = \|\mathbf{i}_k(x, y) - \mathbf{i}_{\text{ref}}(x, y)\| \tag{4}$$

and a subsequent thresholding. For those pixels with a difference exceeding the threshold, a bounding box will be calculated, and its parameters (center, width, height) are combined in a motion measurement vector with the components $\mathbf{z}_{\text{m}} = [x_{\text{cobb,m}}, y_{\text{cobb,m}}, w_{\text{bb,m}}, h_{\text{bb,m}}]^T$.

4 Combining multiple modes

A very interesting aspect of the condensation algorithm is the possibility to rather efficiently integrate the data of several measurements. As mentioned in the introduction, such a combination can make it possible to overcome disadvantages of a single method and to combine the strong points of several methods.

The point where we merged our measurements into the condensation algorithm is the calculation of the weights $\pi_k^{(n)}$ for the sample vectors $\mathbf{s}_k^{(n)}$. Thus, if one has as for example two (normalized) measurement probabilities which are obtained from (2), using different measurement vectors and appropriate measurement matrices, the resulting sample weight is calculated by multiplying them according to the equation

$$p(\mathbf{z}_1, \mathbf{z}_2|\mathbf{s}_k^{(n)}) = p(\mathbf{z}_1|\mathbf{s}_k^{(n)}) \cdot p(\mathbf{z}_2|\mathbf{s}_k^{(n)}) \tag{5}$$

and a subsequent normalization. These modified sample weighting probabilities will have a strong impact on the tracking result, which is now the result of a multimodal fusion of different information channels.

5 Results

Some interesting results of our tracking algorithm are depicted in Fig. 1 and Fig. 2, where the bold white bounding box indicates the expectation value of the samples.

In Fig. 1 an indoor tracking scenario with a panning camera is depicted. Here the major difficulty is that the legs of the person are partially occluded by the desks in the foreground while the person is walking along behind them. Because our P2DHMM was trained only to fully visible persons, it has some problems in this case, and the tracker using only the measurement data of the P2DHMM will fail after a while, as can be seen in the upper row. In the lower row however we see the results after we combined the P2DHMM with a skin finder which is calculating the center of the skin pixels in the upper part of the search region (indicated by the large bounding box) which should be nearly the face of the person. This measurement is indicated by a white cross. As can be seen, now our tracker with combined resources is capable of tracking this sequence. If the tracking process is solely based on the skin finder, it fails as well because this measurement alone is quite unreliable. Thus, both modes support each other in an optimal manner.

In Fig. 2 a typical outdoor surveillance scenario with a non moving background is depicted (data from PETS 2001). For this sequence we used a combination of a P2DHMM and a motion detector. In the upper row we can see a case where the system with the P2DHMM mode alone loses the track after a while (see the third frame in this row), whereas in the lower row it can be seen that after integration of the motion detector mode the system keeps the track. In the last frame in the upper row a detailed result of the motion detector with the detected motion area and its bounding box can be seen. Also here, the use of the motion detector as single measurement mode will fail because other moving objects (see the passing car in the second frame) are severely disturbing this measurement.

6 Conclusion

In this paper we presented a novel approach for a multimodal tracking system based mainly on a variant of the condensation algorithm and a P2DHMM. The architecture of this system has been described and implemented, and some exemplary results have been shown. The major innovation of our approach is the computation of conditional probabilities from the measurement vectors and the probabilistic mode fusion based on these values. Tests have shown that the combination of several tracking modes is a suitable approach to increase the performance of a tracking system in critical scenarios where a single approach alone fails.

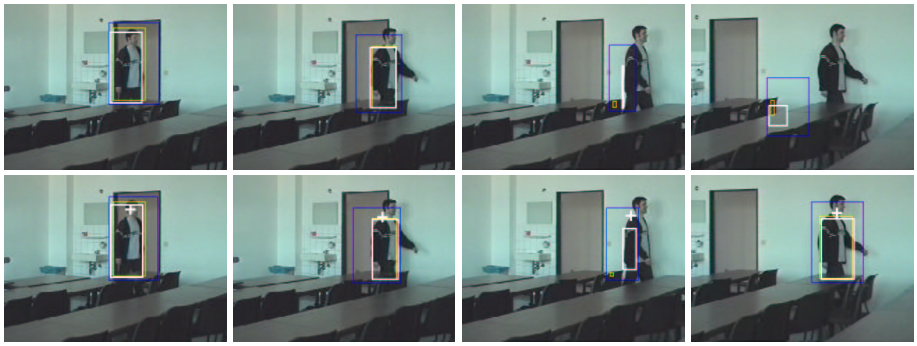


Fig. 1. Tracking results on a difficult indoor sequence with partial occlusion of the lower body. Upper row: Only P2DHMM. Lower row: P2DHMM combined with the skin finder (indicated by a white cross). See text.

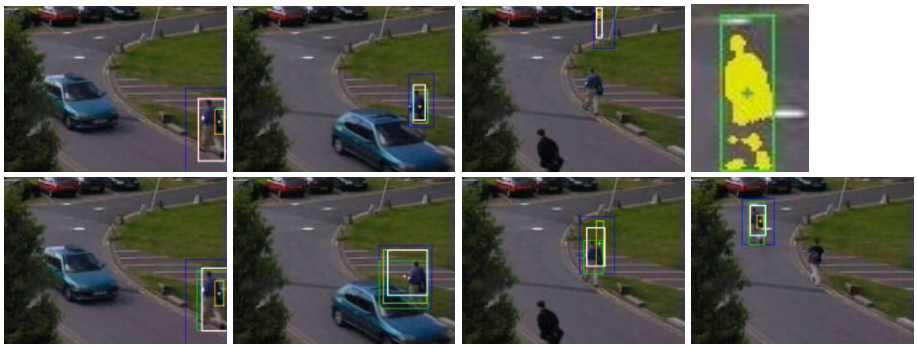


Fig. 2. Tracking results on a realistic outdoor surveillance sequence. Upper row: Only P2DHMM. Lower row: P2DHMM combined with the motion detector (indicated by an additional bounding box). See text.

References

1. H. Breit and G. Rigoll. Improved Person Tracking Using a Combined Pseudo-2D-HMM and Kalman Filter Approach with Automatic Background State Adaptation. In *Proceedings of the ICIP*, Thessaloniki, Greece, Oct. 2001.
2. M. Isard and A. Blake. Condensation – Conditional Density Propagation for Visual Tracking. *International Journal of Computer Vision*, 29(1):5–28, Aug. 1998.
3. L. R. Rabiner and B. H. Huang. An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, Jan. 1986.
4. G. Rigoll, S. Eickeler, and S. Müller. Person Tracking in Real-World Scenarios Using Statistical Methods. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Grenoble, France, Mar. 2000.
5. G. Rigoll, S. Eickeler, and I. K. Yalcin. Performance of the Duisburg Statistical Object Tracker on Test Data for PETS2000. In *IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–7, Grenoble, France, Mar. 2000.
6. K. Schwerdt. *Appearance-Based Video Compression.* PhD thesis, Inst. National Polytechnique de Grenoble, May 2001.

Ideal-observer-model and psychophysical experiments on the role of form information in biologial motion perception

Joachim Lange, Karsten Georg, Markus Lappe

Department of Psychology, Westf. Wilhelms-University Münster, Germany

Abstract. 'Biological motion perception' refers to the impressive ability of human observers to visually identify the motion of humans or animals solely from the moving patterns of a small number of light points attached to the body. Although the first experiments concerning the perception of biological motion already took place in 1973 [1] the perceptual mechanisms are still poorly understood. Based on experiments with a novel biological motion stimulus Beintema and Lappe [2] recently proposed that the perception of biological motion relies more on form than on motion signals. We developed an ideal-observer-model which is based on form information only. In various forced-choice experiments we compared the model's performance with that of human observers in psychophysical studies. The model results showed striking similarities with the data from human subjects. These findings lend additional support to the idea that biological motion perception is based on an analysis of sequential poses each derived from form signals.

1 Introduction

A walking human person produces a highly complex visual motion pattern. However, despite its non-rigidness and its many degrees of freedom this pattern can be recognized by human observers in a fraction of a second. Johansson [1] revealed that this is even true when the visible information is reduced to only a few light points fixed on the joints of the walker. The information transmitted by this 'point-light' display, which is commonly presented as a computer animation [3], can be subdivided into motion and position signals (figure 1a). A single frame of this animation provides form information via the joint positions. A sequence of frames provides motion information via apparent motion signals of the individual points. Since a single frame does not induce the percept of biological motion in naive observers, many studies and models argued that the rapid recognition of biological motion is based on motion signals [1, 4]. Interestingly, however, some patients with lesions in the motion processing areas of the brain are impaired in perception of general aspects of image motion but not in the recognition of biological motion [5, 6].

Beintema and Lappe hypothesized upon these findings that the recognition of biological motion is based on spatiotemporal integration of form information

rather than directly on motion signals [2]. They created a new biological motion stimulus by placing light points at random positions on the extremities rather than on the joints, and then removed local motion signals by jumping points randomly to new positions on the body for each animation frame. Psychophysical

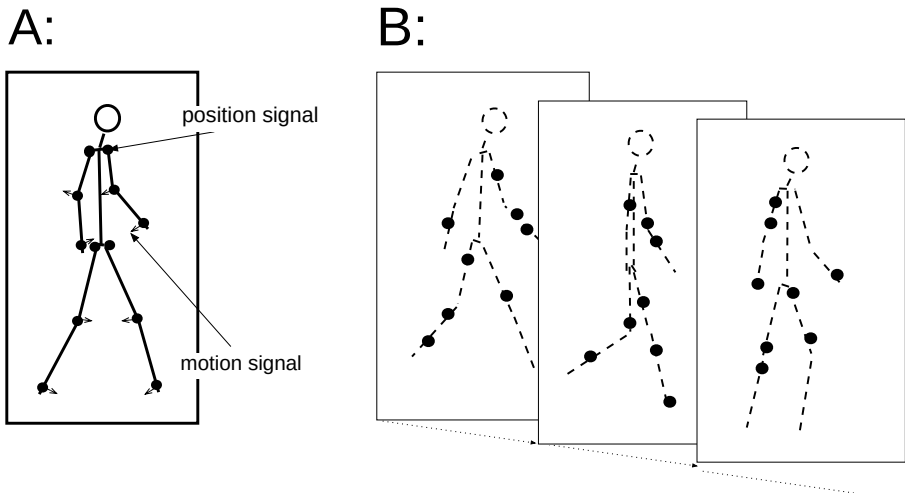


Fig. 1. A: Subdivision of the signals from the walker into position and motion components. B: The single-frame-lifetime (SFL) stimulus consisted of dots that changed their position on the limbs randomly from frame to frame

studies with these 'single-frame-lifetime' (SFL) stimuli showed that biological motion was still perceived from this stimulus, and that two classical 2AFC tasks, direction (SFL-Walker walking either to the right or to the left) and coherence (upper and lower part of the SFL stimulus walking either in the same or in opposite direction) discrimination, could be performed reliably [2]. In the present work, we developed an ideal-observer-model based on position signals in order to obtain a quantitative grasp on the role of position information in the perception of biological motion. We analyzed model behavior and compared it to experimental data.

2 Methods

2.1 Experiments

For the classical biological motion stimulus, we used an algorithm adapted from Cutting [3]. It computes the joint positions for a point-light display (classical walker) giving the impression of a person walking on a treadmill. For the SFL stimulus, the point-light positions were computed to be somewhere between the

joints, the exact placement changing randomly from one frame to the next. The walker subtended 5 by 11 degrees of visual angle and consisted of white dots. Each animation frame was shown for 52ms. The entire stimulus lasted 2.1s. Ref. [2] provides more detailed information on the stimulus.

In each experiment 2-6 observers participated. They watched the walker stimulus on a dark monitor screen and performed one of several discrimination tasks.

2.2 Simulations

Experimental discrimination tasks were recreated in model simulations. The model used an internal standard of a human walker. We recorded the limb movements of 9 human walkers with a motion tracking system (Ascension MotionStar). A step cycle of the average of these walkers was subdivided into 100 temporally equidistant frames acting as the internal model of the limb configurations of a human walker during a step cycle. For every stimulus frame in the experiment simulation, the model computed the mean distances between the dots in the stimulus frame and the limbs for each frame of the internal standard (figure 2b). The decision for every stimulus frame was then based on the set of standard frames with the minimum distance.

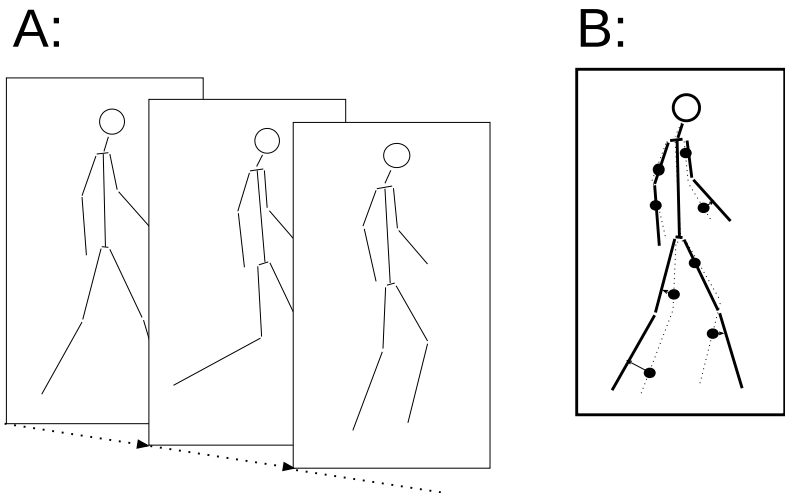


Fig. 2. A: The internal standard consisted of a step cycle of an average human walker subdivided into 100 frames, B: The model's decision is based on linear distance measurements between internal standard and stimulus

In the case of right/left discrimination the model's internal standard consisted of 100 frames of a walker facing and walking to the right and the same number of frames for a walker moving to the left. After the entire stimulus sequence was analyzed, the single answers for each stimulus frame were averaged

to yield an over-all decision. The same approach was taken in the case of coherent/incoherent discrimination, the only difference being that the model's internal walker was subdivided into upper and lower part of the body, a left/right decision was made for each part separately, and then the two decisions were compared for coherence. In both tasks, the model's decisions were therefore based entirely on position information and did not include apparent motion signals between frames.

In the model, we must take into consideration that because of visible persistence [7] for frame durations smaller than 100 ms the number of point-lights perceived at any moment in time is more then the number shown on the display. For instance, for 52 ms frame duration the number of points perceived is about twice the number of dots presented in one frame. To mimic the effect of visible persistence, the model always superimposed any individual frame with the immediately preceding one.

3 Results

3.1 Influence of number of points

As a first quantitative determinant of form information we varied the number of points per frame in several 2AFC tasks. In the direction task, model and human observers had to judge whether the SFL-walker was facing to the right or to the left. In the coherence task they had to discriminate between a coherent and an incoherent walker. A step cycle of the stimulus consisted of 40 frames with a duration of 52 ms (5 monitor refreshes) each. Figure 3a,b shows that the

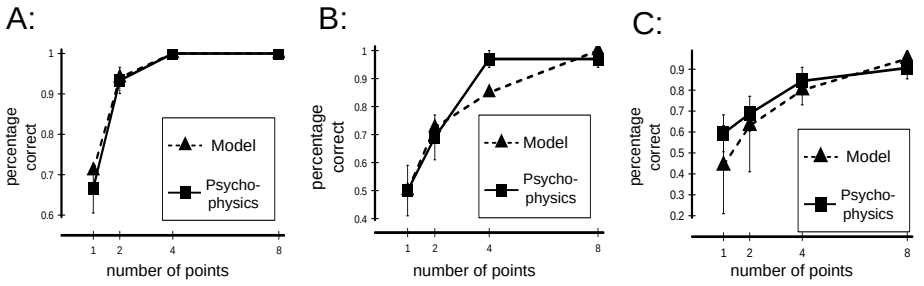


Fig. 3. Comparison of correct answers between model and psychophysical data for A: right/left - , B: coherent/incoherent - and C: forward/backward discrimination

percentage of correct answers increased with rising number of points, both for the model and for the human observers. The similarity between model and human data is surprising as the model does not use any information about the local motion of the points nor about the sequence of the frames. This suggests that the major information used by human observers in the direction and coherence

tasks is frame-by-frame position information, rather than motion signals derived from an analysis of the frame sequence.

We next wished to study a task which cannot rely on form information alone, but which requires sequence analysis. Therefore, we asked observers in a further experiment to discriminate a forward moving display from a backwards moving display. This required the analysis of temporal order over animation frames. The model computed again the distance measures for individual frames but thereafter took the temporal order of the frames into account. Again, performance strongly depended on the number of points per frame (figure 3c). However, the slope was not as steep as for the two previous tasks and performance did not reach 100 percent. Nevertheless, model and psychophysics were again strikingly similar.

3.2 Influence of point lifetime

Beintema and Lappe [2] investigated the potential contribution of local motion signals by prolonging the time over which each light point stayed at one position before jumping to another position (52, 104, 208, or 416 ms) in the direction discrimination task. They argued that if local motion contributes to the perception of biological motion one would expect the percentage of correct answers to increase with prolonged lifetime. But instead of an increase the performance remained constant or showed even a slight decrease with longer lifetimes. Beintema and Lappe speculated that perhaps the reduction in the number of independent position samples that resulted from the increased lifetime led to the decrease in performance.

Model simulations supported this hypothesis (figure 4) as they revealed the same qualitative and quantitative behavior as psychophysical data. This confirms that human observers do not take advantage of additional motion signals. Instead the reduced position information leads to a decline in correct perception rate.

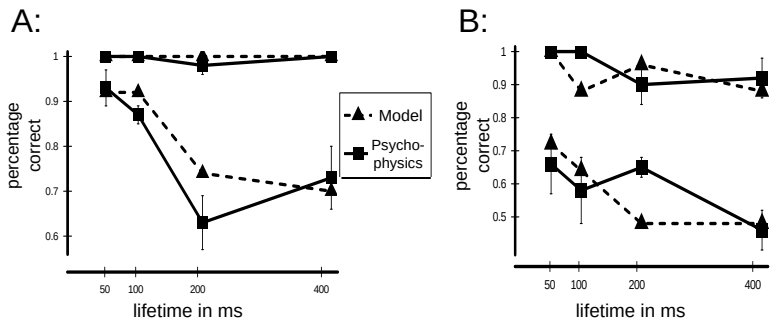


Fig. 4. The influence of lifetime on the percentage of correct answers for A: 8 (two upper curves) and 2 (two lower curves) points and B: for 4 (two upper curves) and 1 (two lower curves) point. Comparison between model and psychophysics

Beintema and Lappe [8] also investigated the potential contribution of local motion signals in the forward/backward discrimination task. In this task, too, prolonged point lifetime did not aid performance. Model simulations showed again similar behavior. No positive influence of prolonged lifetime on the correct answers was observed. This strengthens the conclusion that motion signals do not contribute to performance in this task.

4 Summary and discussion

We investigated the role of position signals in the perception of biological motion using a novel biological motion stimulus that allowed to vary the availability of motion signals. We compared psychophysical studies with an ideal-observer-model that relied only on position information. All experiments revealed striking similarities between model and human data. This suggests that perception is possible from the analysis of form information alone. The model demonstrated that two common psychophysical tasks, direction discrimination and coherence discrimination, could be solved with the same accuracy as human observers without using any motion information. A further task, the discrimination between forward and backward display of a walking person, clearly involved a judgment of motion direction. The model was able to solve this task with the same accuracy as human observers by first analyzing static postures of single frames and then the order of frames in the sequence. Thus, also in this case visual motion signals were not needed.

6 Literatur

1. G. Johansson. Visual perception of biological motion and a model for its analysis. *Percep.Psychophys.*, 14:201–211, 1973.
2. J. A. Beintema and M. Lappe. Perception of biological motion without local image motion. *Proc.Nat.Acad.Sci.USA*, 99:5661–5663, 2002.
3. J. E. Cutting. A program to generate synthetic walkers as dynamic point–light displays. *Behav.Res.Meth.Instrumentation*, 10:91–94, 1978.
4. J. E. Cutting. Coding theory adapted to gait perception. *J.Exp.Psychol.: Hum.Percept.Perform.*, 7:71–87, 1981.
5. L. Vaina, M. Lemay, D. C. Bienfang, A. Y. Choi, and K. Nakayama. Intact biological motion and structure from motion perception in a patient with impaired motion mechanisms: A case study. *Vis.Neurosci.*, 5:353–369, 1990.
6. P. McLeod, W. Dittrich, J. Driver, D. Perrett, and J. Zihl. Preserved and impaired detection of structure from motion by a "motion-blind" patient. *Vis.Cognition*, 3:363–391, 1996.
7. M. Coltheart. Iconic memory and visible persistence. *Percep.Psychophys.*, 27:183–228, 1980.
8. J. A. Beintema and M. Lappe. The role of local position and motion signals in biological motion perception. *Perception*, 30 (suppl.), 2001.

The little difference: Fourier based synthesis of gender-specific biological motion

Nikolaus Troje

Fakultät für Psychologie, Ruhr-Universität-Bochum, 44780 Bochum
troje@uni-bochum.de, www.bml.psy.ruhr-uni-bochum.de

Abstract. A framework is outlined that can be employed to obtain gender and other characteristics of the agent from human motion patterns and subsequently use this information to synthesize motion with particular, well-defined biological and psychological attributes. The proposed model is based on the statistics of a data base of motion capture data. Based on linearization of the motion data, a motion space is defined which is spanned by the first few principal components obtained from the data base of input walkers. Using biological and psychological traits attributed to the input walkers, linear discriminant functions are computed which define vectors in the motion space that generalize the respective trait. These vectors are in turn used to generate walking patterns with the respective properties.

1 Introduction

Biological motion contains plenty of information about identity, personality traits and emotional state of the moving person. The human visual system is extremely sensitive to retrieve such information from motion patterns. We can recognize a familiar person by the way he or she walks and we can attribute gender and age as well as psychological attributes such as personality traits and emotions to an unfamiliar person with motion being the only source of information. We are also extremely sensitive in detecting deviations from natural behaviour. The high degree of perceived realism of modern computer graphics in animated movies and computer games is often disturbed by the fact that the animated movements are perceived to be unnatural. For modern atavars or in the case of virtual replacements of real actors ("virtual stunt men") the observer is not supposed to even realize that the real actor is temporarily replaced by a digital character. To achieve the desired realism, there is considerable demand on methods to synthesize psychologically convincing biological motion.

I want to outline a framework that can be employed to obtain parameterizations of biological or psychological attributes from human motion. Subsequently, I will use this information to synthesize motion with the respective attributes. Gender classification is used as the main example, but I also present examples of how the framework can be applied to other attributes.

The data material to start with is raw motion capture data, i.e. the three-dimensional trajectories of discrete points on a persons body. The primary goal is to transform those data into a representation that would allow us to apply standard methods from

linear statistics and pattern recognition. Such representations have been termed "morphable models" [1-3] in the computer vision community, expressing the fact that the linear transition from one item to a second item of the data set represents a well defined, smooth metamorphosis. Another term that has been used for the same class of models in the context of human face recognition is "correspondence-based representations" [4,5]. This term focuses on the fact that morphable models rely on establishing correspondence between features across the data set resulting in a separation of the overall information into range specific information on the one hand and domain specific information on the other hand [6].

The procedure developed in the present study contains elements of earlier work on parameterizations of animate motion patterns [1,7-10]. Unuma [7] showed that blending between different motions works much better in the frequency domain. At least for periodic motions, such as most locomotion patterns, Fourier decomposition can be used to achieve efficient, low-dimensional, linear decompositions. In fact, decomposing the time series of postures of a single walking person by means of principal component analysis reveals components, which are almost similar to Fourier components [10]. This demonstrates that Fourier decomposition of walking data is nearly optimal in terms of covering a maximum of variance with a minimum of components.

The focus of the current study is to obtain a system that is sensitive enough to extract biologically and psychologically relevant attributes. Based on the linearization of the motion data, a motion space is defined which is spanned by the first few principal components obtained from a set of input walkers. Within this space, linear discriminant functions are computed that generalize the respective trait. Those vectors are in turn used to generate walking patterns with the respective properties in a psychologically convincing manner.

2 Linearization of motion capture data

For the current study, twenty men and twenty women, most of them students and staff of the Psychology department of the Ruhr-University served as models to acquire motion data. A set of 38 retroreflective markers was attached to their body. Participants wore swimming suits and most of the markers were attached directly to the skin. Others, like the ones for the head, the ankles and the wrists were attached to elastic bands and the ones on the feet were taped onto the subjects' shoes.

Participants were then placed on a treadmill and were asked to walk. They could adjust the speed of the treadmill such that they felt most comfortable. To ensure that they did not feel too much under observation and that they did not "perform" in an unnatural manner, we let them walk for at least 5 minutes before we started to record 20 steps (i.e. 10 full gait cycles) from each of them.

Recording was done by means of a motion capture system (Vicon 512, Oxford Metrics). The system tracks the three-dimensional trajectories of the markers with spatial accuracy in the range of 1 mm and a temporal resolution of 120 Hz.

Based on the trajectories of the 38 original markers, we computed the location of "virtual" markers positioned at major joints of the body. The 15 virtual markers used

for all the subsequent computations where located at the ankles, the knees, the hip joints, the wrists, the elbows the shoulder joints, at the center of the pelvis, on the clavicula and in the center of the head.

The walk of an individual subject can be regarded as a time series of postures. Each posture can be described in terms of the position of the 15 markers. Since three coordinates are needed for each position the representation of a single posture is a 45 dimensional vector $p=(m1_x, m1_y, m1_z, m2_x \ldots m15_z)^T$.

Linearization of the data was achieved in two steps. In the first step, the series of postures obtained from a single walker j was decomposed into a second order Fourier expansion:

$$p_j(t) = p_{j,0} + p_{j,1} \sin(\omega_j t) + p_{j,2} \cos(\omega_j t) + p_{j,3} \sin(2\omega_j t) + p_{j,4} \cos(2\omega_j t) + err_j \quad (1)$$

The power carried by the residual term *err* is less than 3% of the power of the input data and we discard it from all further computations. A particular subject's walk is therefore approximated by specifying the average posture $p_{j,0}$, the four characteristic postures $p_{j,1}$, $p_{j,2}$, $p_{j,3}$, and $p_{j,4}$, and the fundamental frequency ω_j. Since each of the components is a 45 dimensional vector, the dimensionality of the model at this stage is 226=5*45+1.

Although this number already reflects a considerable reduction in dimensionality as compared to the raw motion capture data the number of effective degrees of freedom within the database is probably much smaller. For classification purposes it is necessary to reduce the dimensionality of the representation such that the number of dimensions becomes much smaller than the number of items represented in the resulting space.

The advantage of the above representation is, that it provides the possiblity to successfully apply linear operations. Linear combinations of existing walking patterns result in new walking patterns which meaningfully represent the transitions between the constituting patterns [7,10]. We can therefore treat the 226 dimensional vector describing the walk w_j of walker j as a point in a linear space of the same dimension and apply linear methods.

This makes it also possible to use principal components analysis in order to further reduce dimensionality. Applying PCA to the set of walkers W results in a decomposition of each walker into an average walker v_0 and a weighted sum of Eigenwalkers v_k.

$$w_j = v_0 + \Sigma k_{i,j} v_i \quad (2)$$

or in Matrix notation:

$$W = V_0 + VK \quad (3)$$

V_0 denotes a matrix with the average walker v_0 in each column. The matrix V contains the Eigenwalkers as column vectors v_i. Matrix K contains the weights (or the scores) $k_{i,j}$ and is obtained by solving the linear equation system:

$$VK = W - V_0 \quad (4)$$

The variance of the first 15 components sums up to 80% of the overall variance. Truncating the expansion (Eq. 2) after the 15th term thus means loosing 20% of the

overall variance. For all further computations we used a space spanned by just those first 15 Eigenwalkers.

3 Gender discriminant function

Given this relatively low-dimensional linear representation of human walking patterns, we can now construct a linear classifier *c* accounting for gender-specific differences in human walking. This is achieved by finding the best solution (according to a least-square criterium) of the overdetermined linear system

$$cK = r \tag{5}$$

r is the row vector containing 80 values r_j accounting for the desired output of the classifier. r_j equals 1 if walker *j* is male and -1 if the walker is female. *K* is the matrix containing the coefficients of each walker in the 15-dimensional Eigenwalker space. The resulting row vector *c* contains the coefficients of the linear discriminant function best accounting for the gender of the walkers.

The invertibility of the representation can be used to visualize what is happening along this discriminant function by displaying walkers $w_{c,\alpha}$ corresponding to different points along this axis as point-light displays or stick figure animations:

$$w_{c,\alpha} = w_0 + \alpha V c^T \tag{6}$$

As above, w_0 denotes the average walker. The matrix *V* contains the first few Eigenwalkers - one in each column. As α changes from negative to positive values the walker appears to change its gender. On our Web page (*http://www.bml.psy.ruhr-uni-bochum.de/Demos/WDP2002.html*), such animations can be viewed and interactively manipulated by changing the value of α.

We have therefore retrieved a vector *c* that generalizes the attribute “gender” in the obtained motion space. Adding or subtracting this vector from a given walker makes its appearance more male or more female, respectively. The same procedure can be used to extract vectors accounting for other attributes as well. For our database, we registered for every walker a number of easily available attributes such as sex, age and weight. In addition to being able to change the perceived gender of a walker, the above mentioned demonstration also visualizes a dimension obtained from using the weight of the walker to compute a respective discriminant function. Light and heavy walkers show clear differences which are easily extracted by our visual system.

Other attributes, however, are not directly available but have to be determined through psychophysical experiments. In such experiments, observers are presented with displays of the 80 walkers and have to rate them on a 6 point scale with respect to the respective attribute. Here, we report the results of rating two different emotional attributes: happiness vs. sadness, and nervousness vs. relaxedness.

4 Psychophysical determination of emotional attributes

The walking patterns were displayed on a computer monitor as point-light displays subtending 5 deg of visual angle. Each of the 15 markers that were used for the above computation was rendered as a white dot on a black background using orthographic projection from one of three different viewpoints (0 deg = frontal view; 30 deg; 90 deg). The display therefore shows the positions of the major joints of the body changing over time. This results in a vivid percept of a walking human body without providing any information about the person except the one carried by the motion itself [11]. Point-light displays have been widely used in experimental psychology in order to isolate biological motion from other cues about identity, psychological and emotional attributes of a person [12-17, to mention just a few of the classic papers].
A single rating session consisted of 80 trials with each walker shown once for 7 s in a randomized order. All walkers within one session were shown from the same viewpoint. In order to indicate their rating observers had to hit one of 6 buttons displayed on the top of the screen above the point-light display by using the computer mouse. An intertrial interval of 3 s, during which a blank screen was shown, separated the trials. Six observers participated in the experiments. For three observers the most left and right buttons were labeled “happy” and “sad”, respectively. The other three observers were presented with the labels “nervous” and “relaxed”. Each observer carried out three sessions, one for each viewpoint, with short breaks between the sessions. The order of the three sessions was counterbalanced across observers.
The average of the ratings (across the three observers in each group and across the three different viewpoints) was used to form a vector r which, in turn, was used to compute the respective discriminant function c according to Equation 5. The animation at *http://www.bml.psy.ruhr-uni-bochum.de/Demos/WDP2002.html* visualizes the results. Animations both along the happy-sad axis as well as along the nervous-relaxed axis give a clear percept of a change in the respective emotions of the walker.

5 Discussion

Visualizing the respective discriminant functions shows that we have really captured the particular attribute and that the resulting walker vividly changes its characteristics in accordance with the intended characteristic. In all four cases examined so far, changes are a complex composite of structural and dynamic properties of the walker. For instance the exaggerated male walker has wider shoulders than hips whereas in the female walker this ratio reverses. Male walkers display considerable lateral body sway whereas this is not the case for female walkers. Hip motion in male walkers is 180 phase shifted with respect to the hip motion in female walkers. The position of the elbows is very different in male and female walkers. Men tend to hold their elbows away from the body whereas women hold them close to the body. In general, the exaggerated man seems to attempt to occupy much more space than the exaggerated woman -- a display not unique to the human species.

The differences in walking between light and heavy walkers are much harder to describe. Heavy persons have a somewhat smaller gait frequency and vertical movement components seem to be more pronounced in light-weighted walkers as compared to heavy walkers. However, there remains a discrepancy between the clear percept of a change in weight and the ability to identify the sophisticated composite features that communicate this information. The power of the proposed method for generating characteristic motion, however, is that it is not necessary to specify the features that carry the impression of changing biological or emotional attributes explicitely. Instead, we can extract them in terms of the statistical features of a data base that contains variations along the dimensions of interest.

References

1. Giese, M.A., Poggio, T.: Morphable models for the analysis and synthesis of complex motion patterns. International Journal of Computer Vision 38 (2000) 59-73
2. Jones, M.J., Poggio, T.: Multidimensional morphable models - a framework for representing and matching object classes. International Journal of Computer Vision 29 (1999) 107-131
3. Shelton, C.R.: Morphable surface models. International Journal of Computer Vision 38 (2000) 75-91
4. Troje, N.F., Vetter, T.: Representations of human faces. In: Taddei-Ferretti, C., Musio, C.(eds.): Downward processing in the perception representation mechanism. World Scientific, Singapore (1998) 189-205
5. Vetter, T., Troje, N.F.: Separation of texture and shape in images of faces for image coding and synthesis. Journal of the Optical Society of America A 14 (1997) 2152-2161
6. Ramsay, J.O., Silverman, B.W.: Functional data analysis. Springer, New York (1997).
7. Unuma, M., Anjyo, K., Takeuchi, R.: Fourier principles for emotion-based human figure animation. Computer Graphics Proceedings of SIGGRAPH 95 (1995) 91-96
8. Bruderlin, A., Williams, L.: Motion signal processing. Computer Graphics Proceedings of SIGGRAPH 95 (1995) 97-104
9. Witkin, A., Popovic, Z.: Motion warping. Computer Graphics Proceedings of SIGGRAPH 95 (1995) 105-108
10. Troje, N.F. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. Journal of Vision (in press)
11. Johansson, G.: Visual perception of biological motion and a model for its analysis. Perception & Psychophysics 14 (1973) 201-211
12. Dittrich, W.H.: Action categories and the perception of biological motion. Perception 22 (1993) 15-22
13. Mather, G., Murdoch, L.: Gender discrimination in biological motion displays based on dynamic cues. Proceedings of the Royal Society of London Series B 258 (1994) 273-279
14. Kozlowski, L.T., Cutting, J.E.: Recognizing the sex of a walker from a dynamic point-light display. Perception & Psychophysics 21 (1977) 575-580
15. Barclay, C.D., Cutting, J.E., Kozlowski, L.T.: Temporal and spatial factors in gait perception that influence gender recognition. Perception & Psychophysics 23 (1978) 145-152
16. Dittrich, W.H., Troscianko, T., Lea, S., Morgan, D.: Perception of emotion from dynamic point-light displays represented in dance. Perception 25 (1996) 727-738
17. Cutting, J.E., Kozlowski, L.T.: Recognizing friends by their walk: Gait perception without familiarity cues. Bulletin of the Psychonomic Society 9 (1977) 353-356

Gabor-based Feature Point Tracking with Automatically Learned Constraints*

Jan Wieghardt[1], Rolf P. Würtz[2], and Christoph von der Malsburg[2,3]

[1] SIEMENS AG, CT SE 1, Otto-Hahn-Ring 6, D-81730 München
jan.wieghardt@mchp.siemens.de
[2] Institut für Neuroinformatik, Ruhr-Universität Bochum, D-44780 Bochum
rolf.wuertz@neuroinformatik.ruhr-uni-bochum.de
[3] LCBV, University of Southern California, Los Angeles, USA

Abstract. All point tracking mechanisms sometimes fail due to ambiguities in the visual data, a problem which can be alleviated by introducing model knowledge in the form of constraints on groups of feature points. Starting from a point tracking mechanism based on Gabor phases we introduce model constraints, on the one hand by posterior regularization (externally) and on the other hand by incorporating them directly into the tracking mechanism (internally). In the special case of facial feature tracking we show how the necessary model knowledge expressed in the constraints can be learned without explicit user interaction. To this end typical transformations of point groups are learned from noisy but automatically determined correspondences via principal component analysis.

1 Introduction

Tracking feature points reliably through a sequence of images is a much desired skill for all applications where trajectories need to be measured and evaluated. In this context *Gabor wavelets* have turned out to be well suited to determine the disparity between two points from consecutive images [3, 4]. The phase of the complex response to a Gabor filter varies nearly linearly for small translations in the image plane [1], which allows disparity estimation with subpixel accuracy. Another important feature are the multi-scale properties providing a very flexible point description and the ability to robustify disparity estimation over a wide range of scales.

Despite these advantages the tracking of individual feature points using Gabor wavelets still suffers from local image ambiguities like the infamous *aperture problem* that cannot be resolved without taking a larger context into account. Such a context can often be provided by a set of constraints on a whole *group* of points to be tracked. We propose a method which allows to incorporate the constraints directly during disparity estimation. Full details about method and results can be found in [5].

* Funding by European Commission in the Research and Training Network MUHCI (HPRN-CT-2000-00111) and the German Federal Minister for Science and Education under the project LOKI (01 IN 504 E 9) is gratefully acknowledged.

2 Disparity estimation

In the tracking algorithm in [3] the *disparity* of a point from one frame to the next is estimated in terms of phase differences of single *Gabor jets* with amplitudes a and phases ϕ. Extracting two jets at positions $\boldsymbol{x}$ and $\boldsymbol{x}'$, their relative disparity $\boldsymbol{d}$ can be calculated by maximizing their similarity

$$s = \frac{\sum_{\boldsymbol{k}} a_{\boldsymbol{k}}(\boldsymbol{x}) a_{\boldsymbol{k}}(\boldsymbol{x}')(1 - 0.5(\phi_{\boldsymbol{k}}(\boldsymbol{x}) - \phi_{\boldsymbol{k}}(\boldsymbol{x}') - \boldsymbol{k}^T \boldsymbol{d})^2)}{|\boldsymbol{J}(\boldsymbol{x})||\boldsymbol{J}(\boldsymbol{x}')|}. \tag{1}$$

The disparity is first estimated using only the lowest center frequency $\boldsymbol{k}$. Afterwards, in each iteration one additional level is added, and the corresponding phase differences are corrected modulo 2π. Thus, the lower frequencies can resolve the natural ambiguity modulo the wavelength for the higher frequencies. In case the estimated intermediate disparity exceeds twice the actual width of the Gabor function on the next higher frequency level, the process is terminated.

3 Tracking individual feature points

A tracking algorithm can be based on this disparity estimation, by executing the following steps for each frame (the parameter α can be adjusted to the expected variability of the visual features during tracking).

1. Extract jets $\boldsymbol{J}_i(\boldsymbol{x}_1(t_i)), \ldots, \boldsymbol{J}_i(\boldsymbol{x}_m(t_i))$ at current positions in frame I_i.
2. Update model jets $\boldsymbol{J}_i^{\text{model}}(\boldsymbol{x}_n(t_i)) = (1-\alpha)\,\boldsymbol{J}_{i-1}^{\text{model}}(\boldsymbol{x}_n(t_{i-1})) + \alpha \boldsymbol{J}_i(\boldsymbol{x}_n(t_i))$.
3. Calculate disparity to the jets extracted from the next image I_{i+1} at the same image-coordinates $\boldsymbol{d}_n = \boldsymbol{d}_n\left(\boldsymbol{J}_i^{\text{model}}(\boldsymbol{x}_n(t_i)), \boldsymbol{J}_{i+1}(\boldsymbol{x}_n(t_i))\right)$.
4. Calculate new positions in image I_{i+1}: $\boldsymbol{x}_n(t_{i+1}) = \boldsymbol{x}_n(t_i) + \boldsymbol{d}_n$.

4 Tracking constrained groups of points

Constraints for the disparities $\boldsymbol{d}_n$ of the points n can only come from a *parameterized model* of the possible variations. They take the general form

$$\boldsymbol{d}_n - \boldsymbol{f}_n(\boldsymbol{\epsilon}) = 0. \tag{2}$$

In this situation $\boldsymbol{f}_n$ is a model of the possible group motion and $\boldsymbol{\epsilon}$ are the model-parameters. E.g., if only image plane rotations are possible, $\boldsymbol{\epsilon}$ would contain the center and angle of the rotation, and $\boldsymbol{f}_n(\boldsymbol{\epsilon})$ the resulting displacement of point n. In practice, the equality is relaxed to a minimization of the norm of the left hand side of (2).

These constraints can be incorporated by first estimating the disparities assuming all nodes to be mutually independent and then calculating the constrained disparity configuration that is closest, in a least square sense, to the estimated disparities. The disparities are subsequently changed to those given by the constrained configuration. This method, which we call *external constraints*,

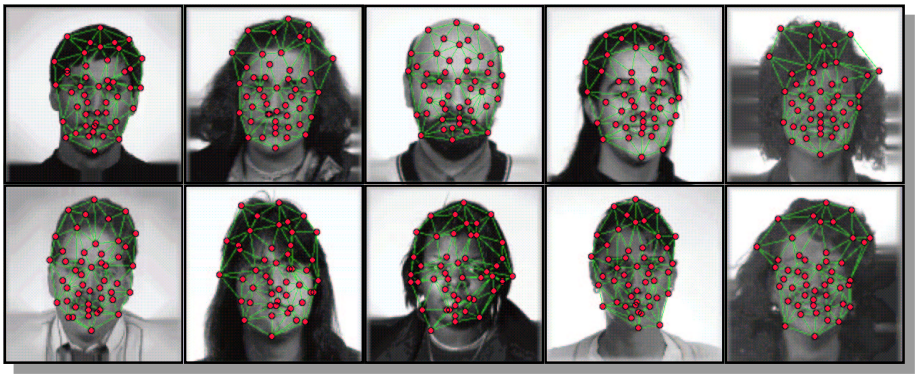


Fig. 1. Examples of automatically labeled faces: Displayed are 10 arbitrarily chosen examples of a set of approximately 1000 images. The retrieved correspondences are displayed by superimposing the *bunch graph*.

has serious drawbacks, as the separation of model knowledge and motion estimation forces decisions while estimating the initial disparities, even if the available image information is inadequate. This can cause small errors to accumulate and the total tracking result to deteriorate.

A better way is integrating the the model knowledge directly into motion estimation. Substituting constraints in the form of equation (2) into the phase-based disparity estimation of equation (1), the constrained disparities can be found by maximizing

$$s(\boldsymbol{\epsilon}) = \sum_n \frac{\sum_{\boldsymbol{k}} a_{\boldsymbol{k}}(\boldsymbol{x}_n) a_{\boldsymbol{k}}(\boldsymbol{x}'_n) \left(1 - 0.5 \left(\phi_{\boldsymbol{k}}(\boldsymbol{x}_n) - \phi_{\boldsymbol{k}}(\boldsymbol{x}'_n) - \boldsymbol{k}^T \boldsymbol{f}_n(\boldsymbol{\epsilon})\right)^2\right)}{|\boldsymbol{J}(\boldsymbol{x}_n)||\boldsymbol{J}(\boldsymbol{x}'_n)|}. \quad (3)$$

Applying a first order Taylor expansion and maximization in terms of $\Delta\boldsymbol{\epsilon}$ yields a linear equation system for $\Delta\boldsymbol{\epsilon}$, which can be solved during the coarse-to fine tracking described above. We term this use of model constraints *internal*.

5 Learning constraints from example data

Having established the need for constraints and a good way to apply them during tracking the question remains of how the correct constraints for an object class can be found. An analytical description will only be feasible in the simplest of cases, and it is desirable to learn the constraints from example images. We demonstrate a solution to this problem on face tracking. We match a bunch graph [6] onto a large set of more or less frontal faces. The resulting *correspondence fields* are converted into vectors and subjected to *Principal Component Analysis (PCA)* in a way similar to [2].

PCA yields the mean deformation and the deformations with the largest variation in the dataset. The first 6 are visualized in figure 2. As it turns out the

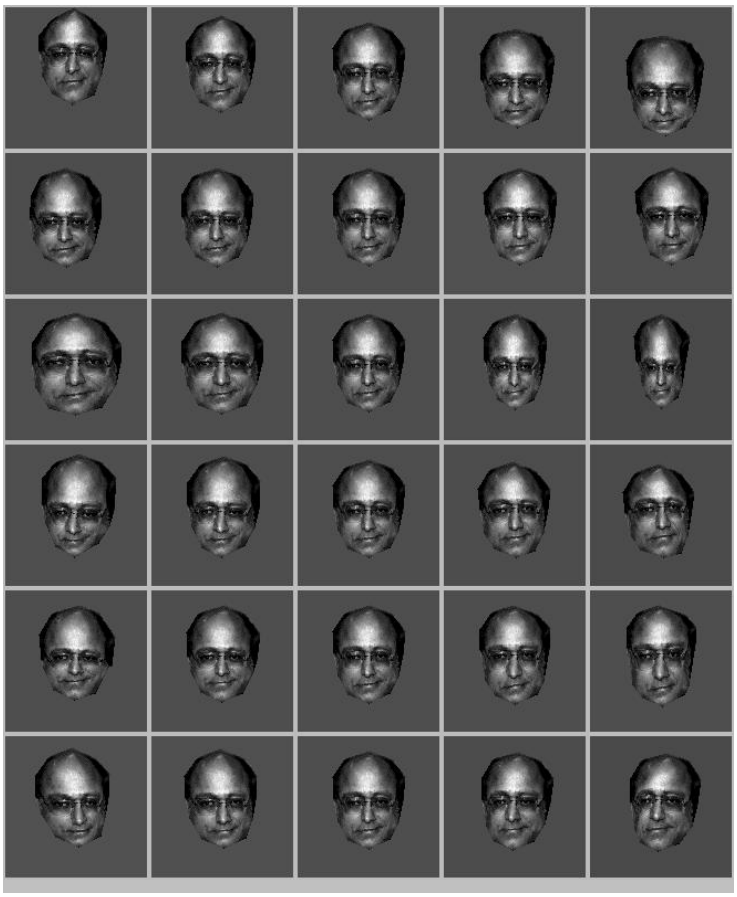
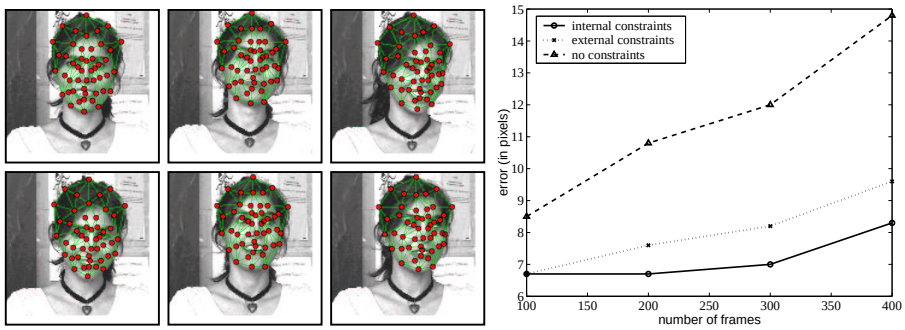


Fig. 2. Textured principal components of correspondence fields: The principal components $\boldsymbol{P}_1$ through $\boldsymbol{P}_6$ (top to bottom) of the feature point locations are illustrated here in terms of the mapping they perform on the standard gray value image shown in the central column. Each row shows the deformation from the mean along one principal component by -4,-2,0,2 and 4 standard deviations, respectively.

principal components are readily interpretable. They code transformations that are easily identified and named by visual inspection. The first one is a mixture of vertical translation and tilt, the second is horizontal translation, the remaining four contain scaling and rotation in depth. This is remarkable for several reasons. First, the results are based on a noisy database of automatically resolved correspondences. Although the database contained a lot of different individuals and was restricted to approximately frontal pose, the inter-individual variations (such as, e.g., jaw size or eye distance) are not dominant. The main variations seem to stem from geometrical variations. The only inter-individual variation visible in the first six components is expressed in the independence of scaling in x- and



(a) Example sequence (b) Tracking error over time

Fig. 3. Tracking of faces: (a) shows the point positions on selected frames of a sequence, (b) the tracking error over time for internal, external, and no constraints, respectively. The constraints were derived from the first six principal components

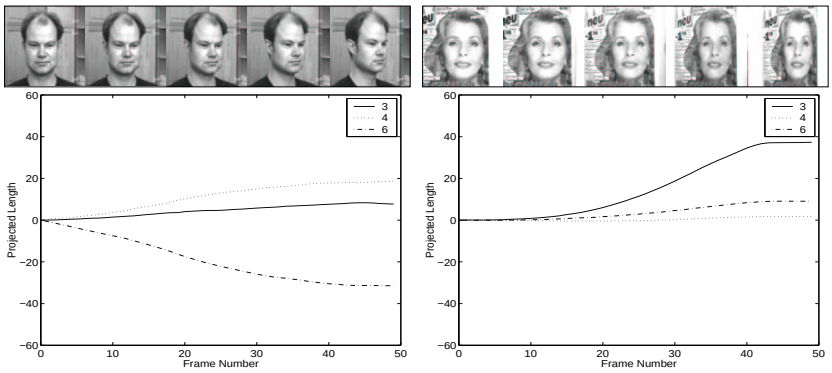
y-direction ($\boldsymbol{P}_3$ and $\boldsymbol{P}_5$), which might be attributed to different head shapes. Although no explicit knowledge about the three-dimensional transformations of rigid objects went into the constraint construction, their main properties were captured. Moreover, the degrees of freedom are nicely separated in an intuitive fashion.

An accurate model of the group motion of the selected feature points can thus be derived by assuming that the whole motion is restricted (or close) to the space spanned by the first principal components $\boldsymbol{P}_1$ through $\boldsymbol{P}_6$. Thus, the projection onto these components can serve directly as model parameters $\boldsymbol{\epsilon}$.

6 Results

Although the correspondences derived from bunch graph matching are far from perfect, the components with the highest eigenvalues seem to capture the major transformations that a face undergoes (see figure 2). They can directly serve as constraints and result in improved tracking performance. The results of three tracking procedures, namely unconstrained tracking, tracking with external constraints and the method proposed here using internal constraints are compared in figure 3(b) and clearly show the superiority of the latter.

Furthermore, the same constraints can be used to give rough pose information and distinguish 3D-motion of a true face from a rotated image of a face. It is remarkable how well the model captures the transformations of a moving face although no image sequences were provided when deriving the model. If the model parameters estimated by projecting the flow fields onto the first PCs are plotted over time for a sequence showing a moving head, as it was done in figure 4, it can be clearly seen that the derived motion model can be used for more than



(a) Head rotated by 30° (b) Journal cover rotated by 45°

Fig. 4. Principal components under rotation in depth: Shown are the projections of the correspondence fields on $\boldsymbol{P}_3$, $\boldsymbol{P}_4$, and $\boldsymbol{P}_6$, respectively, component as functions of the frame number for a real head (a) and a flat photograph of a head (c) monotonously rotating in depth. It can be clearly seen that $\boldsymbol{P}_4$ and $\boldsymbol{P}_6$ are closely related to a head's rotation in depth and its three-dimensional structure.

constraining the tracking. The transformation properties of faces, especially their behavior under rotation in depth, are so well captured that the model parameters themselves can be exploited to yield at least a qualitative pose estimation. The experiment with the journal cover shows that the resulting horizontal scaling can be clearly separated from the 3-D rotation of a real face.

References

1. D. J. Fleet and A. Jepson. Computation of component image velocity from local phase infor mation. *International Journal of Computer Vision*, 5(1):77–104, 1990.
2. A. Lanitis, C. Taylor, and T. Cootes. Automatic interpretation and coding od face images using flexible models. *IEEE Trans. PAMI*, 19(7):743–756, July 1997.
3. T. Maurer and C. von der Malsburg. Tracking and learning graphs and pose on image sequences of faces. In I. Essa, editor, *Proc. 2nd AFGR*, pages 176–181. IEEE Computer Society Press, 1996.
4. S. J. McKenna, S. Gong, R. P. Würtz, J. Tanner, and D. Banin. Tracking facial feature points with Gabor wavelets and shape models. In J. Bigün, G. Chollet, and G. Borgefors, editors, *Proc. 1st AVBPA*, pages 35–42. Springer, 1997.
5. J. Wieghardt. *Learning the Topology of Views: From Images to Objects.* Shaker Verlag, Aachen, 2001.
6. L. Wiskott, J.-M. Fellous, N. Krüger, and C. v.d. Malsburg. Face recognition by elastic graph matching. *IEEE Trans. PAMI*, 19(7):775–779, 1997.

Modeling of movement sequences based on hierarchical spatial-temporal correspondence of movement primitives

Winfried Ilg and Martin Giese

Laboratory for Action, Representation and Learning
Department for Cognitive Neurology, University Clinic Tübingen, Germany
{wilg,giese}@tuebingen.mpg.de

Abstract. In this paper we present an approach for the modeling complex movement sequences. Based on the method of Spatio-Temporal Mophable Models (STMMs) [7] we derive a new hierarchical algorithm that, in a first step, identifies movement elements in the complex movement sequence based on characteristic events, and in a second step quantifies these movement primitives by approximation through linear combinations of learned example movement trajectories. The proposed algorithm is used to segment and to morph sequences of karate movements of different people and different styles.

1 Introduction

The analysis of complex movements is an important problem for many technical applications such as computer vision, computer graphics, sports and medicine. For several applications it is crucial to model movements with different styles. One method that seems to be very suitable to synthesize movements with different styles is the linear combination of movement examples. Such linear combinations can be defined efficiently on the basis of spatio-temporal correspondence. The technique of Spatio-Temporal Morphable Models (STMMs) defines linear combinations by weighted summation of spatial and temporal displacement fields that morph the combined prototypical movement into a reference pattern. This method has been successfully applied for the generation of complex movements in computer graphics (motion morphing), as well as for the recognition of movements and movement styles from trajectories in computer vision [7].

In this paper, we extend the basic STMM algorithm by introducing a second hierarchy level that represents motion primitives. Such primitives correspond to parts of the approximated trajectories, e.g. individual facial expressions or techniques in a sequence of karate movements. These movement primitives are then modeled using STMMs by linearly combining example movements. This makes it possible to learn generative models for sequences of movements with different styles. The extraction of movement primitives is based on simple invariant features that are used to detect key events that mark the transitions between different primitives. Sequences of such key events are then detected by matching

them to a learned example sequence. This matching is based on standard sequence alignment methods that are based on dynamic programming. We apply this hierarchical algorithm to model and synthesizes complex karate movements. In particular, we show that movement primitives from different actors and with different styles can be generated and recombined to longer naturally-looking movement sequences.

2 Algorithm

2.1 Morphable Models as Movement Primitives

The technique of *spatio-temporal morphable models* [6, 7] is based on linearly combining the movement trajectories of prototypical motion patterns in space-time. Linear combinations of movement patterns are defined on the basis of spatio-temporal correspondences that are computed by dynamic programming [2]. Complex movement patterns can be characterized by trajectories of feature points. The trajectories of the prototypical movement pattern n can be characterized by the time-dependent vector $\boldsymbol{\zeta}_n(t)$. The correspondence field between two trajectories $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$ is defined by the spatial shifts $\boldsymbol{\xi}(t)$ and the temporal shifts $\tau(t)$ that transform the first trajectory into the second. The transformation is specified mathematically by the equation:

$$\boldsymbol{\zeta}_2(t) = \boldsymbol{\zeta}_1(t + \tau(t)) + \boldsymbol{\xi}(t) \tag{1}$$

By linear combination of spatial and temporal shifts the spatio-temporal morphable model allows to interpolate smoothly between motion patterns with significantly different spatial structure, but also between patterns that differ with respect to their timing.

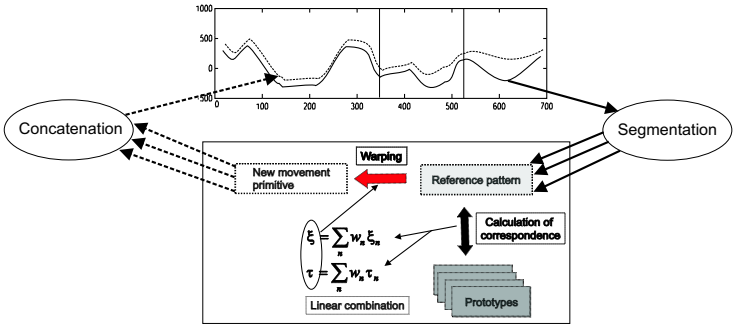


Fig. 1. Schematic description of the algorithm to analyze and synthesize complex movement sequences. In the first step the sequence is decomposed into movement primitives. These movement primitives can be analyzed and changed in style defining linear combinations of prototypes with different linear weight combinations. Afterward the individual movement primitives are concatenated again into one movement sequence. With this technique we are able to generate sequences containing different styles.

The correspondence shifts $\boldsymbol{\xi}(t)$ and $\tau(t)$ are calculated by solving an optimization problem that minimizes the spatial and temporal shifts under the constraint that the temporal shifts define a new time variable that is always monotonically increasing. For further details about the underlying algorithm we refer to [6, 7].

Figure 1 shows schematically the proceeding for generating linear combinations of spatio-temporal patterns for complex movement sequences.

2.2 Representation of Key Features for Movement Primitives

For the identification of movement primitives within a complex movement sequences it is necessary to identify characteristic features that are suitable for a robust and fast segmentation. Different features have been proposed in the literature [4][3]. The key features of our algorithm are based on zeros of the velocity in few "characteristic coordinates" of the trajectory $\boldsymbol{\zeta}(t)$. These features provide a coarse description of the spatio-temporal characteristics of trajectory segments that can be matched efficiently in order to establish correspondence between the learned movement primitives and new trajectories. For the matching process that is based on dynamic programming (see section **??**) we represent the features by discrete events. Let m be the number of the motion primitive and r the number of characteristic coordinates of the trajectory. Let $\boldsymbol{\kappa}(t)$ be the "reduced trajectory" of the characteristic coordinates that has the values $\boldsymbol{\kappa}_i^m$ at the velocity zeros. The movement primitive is then characterized by the vector differences $\Delta\boldsymbol{\kappa}_i^m = \boldsymbol{\kappa}_i^m - \boldsymbol{\kappa}_{i-1}^m$ of subsequent velocity zeros (see figure 2). A formal description of the algorithm can be found in [8].

3 Experiments

We demonstrate the function of the algorithm by modeling movement sequences from material arts. Using a commercial motion capture system (VICON) with 6 cameras and a sampling frequency of 120 Hz we have captured several movement sequences representing a "Kata" from karate from two actors. The first actor was a third degree black belt in Jujitsu, the second actor had the 1. Kyu degree in karate (Shotokan). Both actors executed the same movement sequence but due to differences of the techniques between different schools of martial arts with different styles. In addition both actors also tried to simulate different skill levels, e.g. by mimicking a yellow belt. Three sequences of actor 1 have been segmented manually resulting in six movement primitives, which served as prototypes to define the morphable models of the first actor (see figure 3). Based on the 6 morphable models prototypical representations with key features for the automatic identification of the movement primitives were generated in the way described in section 2.1. The "reduced trajectories" $\boldsymbol{\kappa}(t)$ consist of the coordinates of the markers on both hands.

3.1 Automatic Identification of Movement Primitives

Figure 4 shows the results for the identification procedure for one sequence of actor 2. The automatic segmentation was successful on for all 16 sequences recorded from both actors. Figure 3 shows a morph that was created based on the automatically identified primitives.

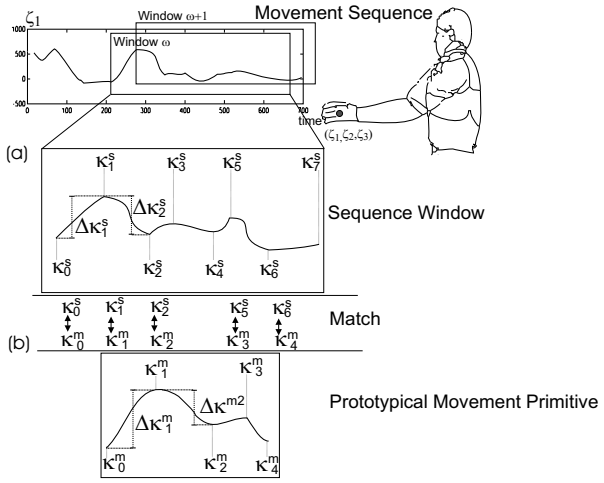


Fig. 2. Illustration of the method for the automatic identification of movement primitives: (a) In a first step all key features $\boldsymbol{\kappa}_i^s$ are determined. (b) Sequences of key features from the sequences (s) are matched with sequences of key features from the prototypical movement primitives (m) using dynamic programming. A search window is moved over the sequence. The length of the window is two times the number of key features of the learned motor primitive. The best matching trajectory segment is defined by the sequence of feature vectors that minimizes $\sum_j ||\Delta\boldsymbol{\kappa}_i^s - \Delta\boldsymbol{\kappa}_j^m||$ over all matched key features. With this method a spatio-temporal correspondence at a coarse level can be established.

3.2 Morphing between different Actors

Based on the movement primitives identified by automatic segmentation morphs between movements of the two actors were realized. The individual movement primitives were morphed and afterward concatenated into a longer sequence. The details of this procedure are described in [5]. Figure 3 shows snapshots from a morphed motion sequence, which corresponds to the "average" of the two original sequences. This sequence looks very natural and shows no artifacts at the margins between the individual movement primitives. In cases, where the styles of both actors are different, the morph generates a realistic movement that interpolates between the styles of the 2 actors original movements. Our technique is thus suitable to generate morphs that cover a continuous spectrum of styles between the actors[1].

4 Discussion

For the Karate data our algorithm successfully morphs between the movements of the same, and of different actors without visible artifacts. In particular the transitions between the individual segments are invisible. The method allows the synthesis of the same Kata with different constant styles, or styles that vary

[1] Movies of the karate animations are provided on the web site http://www.uni-tuebingen.de/uni/knv/arl/arl-demos.html

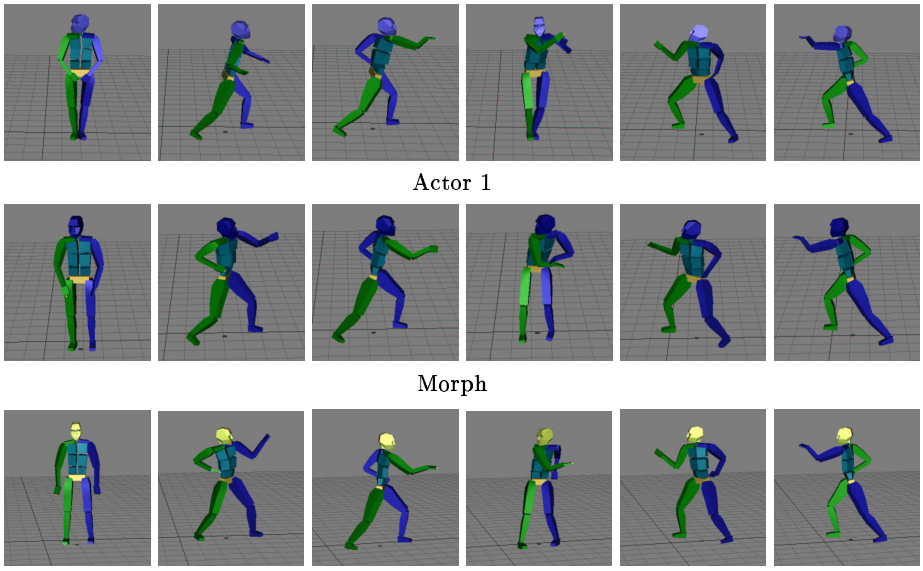


Fig. 3. Snapshots from a sequence of karate movements executed by two actors and a motion morph. The pictures show the initial posture at the beginning and the end postures of the movement primitives 1-5. The end posture is similar to the initial posture. The morphed sequence looks natural and there are no artifacts at the transitions between the 6 movement primitives. Especially interesting is the comparison between the different karate styles of the actors that becomes obvious in the third movement primitive (4th column). Actor 1 is doing a small side step with the left foot for turning. Instead of this, actor 2 turns without sidestep. The morph executes a realistic movement that interpolates the two actors.

over the movement sequence. We were also able to create exaggerations of the individual styles [5].

Interestingly, the algorithm even in the present very elementary form does not lead to the artifact that the feet are slide on the ground plane. This seems to be understandable because correct correspondence between the prototypical movements automatically implies that these constraints are fulfilled by the morphs. However, we expect that morphing between very dissimilar movements in uneven terrains might require to introduce a special handling of such constraints.

Several other approaches rely on statistical methods like hidden markov models to perform a segmentation of movement trajectories [1] [4] [3]. The reason why we prefer dynamic programming, is that our algorithm is also designed for the quantitative analysis of patients with rare movement disorders [9].This requires algorithms, that contrary to most HHM-based methods work efficiently with very small amounts of data. Our method has also been applied successfully to face movements [5]. We think that the method is interesting for a number of applications. Beyond obvious applications in computer graphics and the quantification of movements in sports, we plan to apply the proposed method for

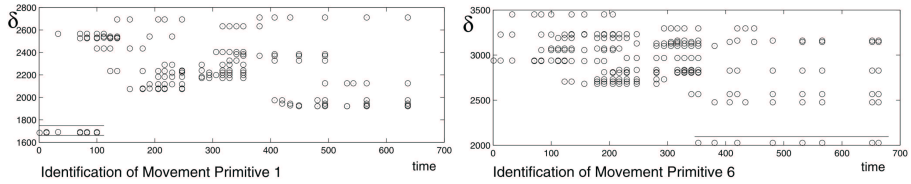


Fig. 4. Results of the automatic segmentation of one movement sequence of actor 2 based on the prototypical movement primitives of actor 1. As an example, the identification of the primitives 1 and 6 is shown. The diagrams show the distance measure δ for different matches of the corresponding movement primitive over the whole sequence. The circles mark the time of the matched key feature κ_i^m in the sequence. Each match of a whole movement primitive is illustrated by a row of circles with the same δ. The number of circles corresponds to the number of key features of the movement primitive (in both diagrams two examples are indicated). Both movement primitives are correctly identified by a minimum in the δ-function.

the generation of stimuli for psychophysical experiments in order to test the recognition of movement sequences in humans.

Acknowledgments

This work is supported from the Deutsche Volkswagenstiftung. We thank H.P. Thier, H.H. Bülthoff and the Max Planck Institute for Biological Cybernetics for additional support.

References

1. M. Brand. Style machines. In *SIGGRAPH*, 2000.
2. A. Bruderlin and L. Williams. Motion signal processing. In *SIGGRAPH*, pages 97–104, 1995.
3. T. Caelli, A. McCabe, and G. Binsted. On learning the shape of complex actions. In *International Workshop on Visual Form*, pages 24–39, 2001.
4. A. Galata, N. Johnson, and D. Hogg. Learning variable lenth markov models of behavior. *Journal of Computer Vision and Image Understanding*, 81:398–413, 1999.
5. M.A. Giese, B. Knappmeyer, and H.H. Bülthoff. Automatic synthesis of sequences of human movements by linear combination of learned example patterns. In *Workshop on Biologically Motivated Computer Vision*, 2002. accepted.
6. M.A. Giese and T. Poggio. Synthesis and recognition of biological motion pattern based on linear superposition of prototypical motion sequences. In *Proceedings of IEEE MVIEW 99 Symposium at CVPR, Fort Collins*, pages 73–80, 1999.
7. M.A. Giese and T. Poggio. Morphable models for the analysis and synthesis of complex motion patterns. *International Journal of Computer Vision*, 38(1):59–73, 2000.
8. W. Ilg and M.A. Giese. Modeling of movement sequences based on hierarchical spatial-temporal correspondence of movement primitives. In *Workshop on Biologically Motivated Computer Vision*, 2002. accepted.
9. W. Ilg, M.A. Giese, H. Golla, and H.P. Thier. Quantitative movement analysis based on hierarchical spatial temporal correspondence of movement primitives. In *11th Annual Meeting of the European Society for Movement Analysis in Adults and Children*, 2002.

Learning of the discrimination of artificial complex biological motion

J. Jastorff *, Z. Kourtzi # and M.A. Giese *

* Group for Action Representation and Learning
Department for Cognitive Neurology, University Clinic Tübingen
Max–Planck–Institute for Biological Cybernetics, Tübingen
{jan.jastorff, zoe.kourtzi, martin.giese}@tuebingen.mpg.de

Abstract. Psychophysical and neurophysiological studies suggest that human body motion presented as point light displays can be readily recognized. So far it has not been investigated whether the recognition takes place on the basis of stored innate templates or if we are able to learn the discrimination of totally new complex motion patterns. To address this question we generated novel artificial biological movement patterns by linearly combining the trajectories of prototypical natural movements in space-time that were recoded by motion capturing. After training the subjects we found a significant improvement in discrimination performance for all stimuli. Our results show that humans are able to learn to recognize completely novel biological motion patterns and that it is possible to learn the discrimination between these artificial stimuli. This suggests that we do not purely rely on stored innate templates.

Introduction

The term biological motion is often used by researchers studying the patterns of movement generated by moving animals and humans. By far the most frequently studied biological motion is human gait. Johansson was the first researcher to investigate systematically the perceptual sensitivity of the visual system for the recognition of biological motion [1]. He developed a technique that minimizes form information by presenting only the joint positions of moving actors as illuminated dots. Psychophysical and neurophysiological studies suggest that human body motion presented in form of such point light displays can be readily recognized. Point light stimuli are sufficient to allow a discrimination of the type of action and even the gender and other details of the walker [2]. However, the perceptual impression of a walker usually breaks down, if the dots are presented as stationary pattern [1]. Nevertheless, the small amount of motion information provided by a few subsequent stimulus frames is sufficient for the visual system to organize the elements into a coherent percept of articulated motion.

Sensitivity to biological motion stimuli arises early in the human development. Sixteen - week - old infants already prefer a point light walker display over the same display rotated by 180 degrees, or dynamic noise stimuli [3].

Additionally it has been shown, that by the age of 3 - 5 months, infants are able to discriminate between a locally rigid point-light walker display and a similar display, in which the local rigidity between the dots is perturbed [4].

The sensitivity of infants regarding biological motion displays, has motivated the hypothesis that perception of biological motion might be an innate capacity of the visual system rather than acquired by learning through experience [3].

In this study we try to investigate whether the recognition and discrimination of biological motion patterns takes place on the basis of stored innate templates, or if humans are able to learn to discriminate completely new complex motion patterns. To address this question we created novel artificial stimuli by motion morphing. This technique allows us to create novel motion patterns that are embeded in a metric space that allows to quantify the spatio-temporal similarity between the morphs.

Methods

Stimulus generation

Stimuli were generated by tracking biological motion from video sequences showing locomotion patterns (walking, running, marching), different types of physical exercises (aerobics, boxing) and martial arts techniques. Twenty one different prototypical motion patterns were recorded.

Motion tracking

The movement patterns were filmed using a Kodak VX 1000 camera with the actor facing and moving on a line orthogonal to the view direction of the camera. All movements were executed periodically, but only a single cycle of the movements was used for motion morphing.

To track the trajectories of the movements, first the translation of the whole body was subtracted from the video sequence by hand-marking the hip position in a number of frames and fitting the translation of the hip by a linear function of time. When the fitted translation was subtracted from the sequence, the resulting movement looked like a person performing the movements on a tread mill. Twelve feature points were tracked manually. These were the head, shoulders, elbows, wrists, hip, knees and ankles.

The tracked trajectories were time-normalized and smoothed by fitting with a second order Fourier series. Afterwards they were used as prototypes for the motion morphing.

Motion morphing

To create morphs between different forms of biological motion we used the technique of *spatio-temporal morphable models* [5]. This method makes it possible to generate new trajectories by linearly combining the movement trajectories of prototypical motion patterns in space-time. Linear combinations of movement patterns are defined on the basis of spatio-temporal correspondences. Complex motion patterns can be characterized by trajectories of feature points, in our case the 2D coordinates of the joints of the moving figure. With the help of this method it becomes possible to interpolate smoothly between motion patterns with significantly different spatial structure and also between patterns that differ with respect to their timing. The twenty one recorded motion patterns were divided into seven groups, each of them containing three prototypical movements. For each group the stimulus trajectories were generated by motion morphing.

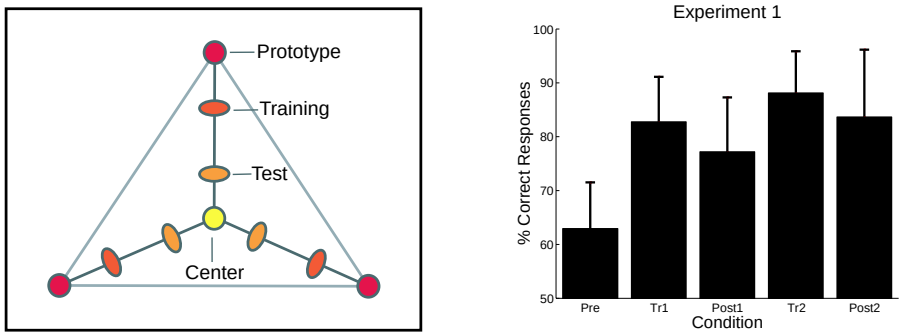


Figure 1: (a) Graphical illustration of the pattern space. The positions of the points in the triangle symbolize the relative contributions of the three prototypes. The distances of the points from the prototypes are related to the weights of the prototypes in the linear combination. (b) Results of experiment 1. Shown are the mean percentage of correct responses (+ s.e.m.) for the five blocks.

Weights define the contributions of the individual prototypes to the linear combination. By adjusting the weights, it becomes possible to create a morph that resembles more or less the individual prototypes. Hence, the class of generated biological movements is equipped with a metric for the spatio-temporal similarity of the patterns.

Separately for each of the seven groups we generated morphs from the three prototypical patterns within this group. The prototypical patterns were arranged in such a way, that the morphs of three of the groups resembled natural human movements, while the other four looked quite unnaturally. The weight of the first prototype was always chosen from the set of [0.33, 0.45, 0.53, 0.56, 0.6, 0.7, 0.8, 0.9, 0.95]. The remaining weights were equal to 1/2 (1 - weight of the first prototype), because the sum of all weights for each morph was restricted to one.

It has been shown, that stimuli created with the help of this technique allow for smooth and continuous variation of the categorization probabilities with the weights of the prototypes in the morph [6].

Stimulus presentation

The biological motion stimuli were generated with an Apple Macintosh G4 computer and displayed on a Sony color monitor (75 Hz framerate; 1024x768 pixels resolution). The monitor was viewed binocularly. The locations of the points in the display were at the ankles, knees, elbows and wrists. Because the locomotion patterns were presented from a side view, only one dot was presented at the hip as well as the shoulder. The 10 black stimulus dots had a diameter of 0.5 degrees of visual angle and were presented within a virtual window, centered on the screen of the monitor on a white background. The size of the whole figure was about 5 x 10 degrees and the center position within the virtual window was randomized uniformly within an interval of $\pm$ 2 degrees horizontally and vertically.

Procedure

Subjects watched the computer screen from a distance of 40 cm. They were briefed about the experimental procedure and were given the opportunity to

practice for seven trials showing one example stimulus from each of the seven groups. In all the trials, morphs with equal weights of the prototypes (center stimuli) were compared to morphs with non-equal weights (off center stimuli) in a pair comparison paradigm. First a single center stimulus was presented for six movement cycles, followed by the same center stimulus next to an off center stimulus generated from the same triple of prototypes, again for six cycles. In a two alternative forced choice test, the subjects were asked to choose the stimulus that was identical to the center stimulus that had been presented before.

The experiment consisted of test blocks and training blocks. In the test blocks, each of the seven groups was presented three times, while in the training blocks every group was shown eight times in random order. The prototype that was contributing the most to the off center stimulus was randomly chosen in all cases. As can be seen in figure 1a), the weights of the off center stimuli were different for test and training. Therefore the test patterns were more similar to the center stimuli than the training ones. In preceding piloting experiments the individual weights for the seven groups were calibrated to assure that the difficulty level within test trials and within training trials was similar. During the training the subjects received feedback after each trial whether their response was correct or not. In the test blocks no feedback was provided.

The experiment consisted of three test blocks intersected by two blocks of training. After the experiment the subjects were asked to categorize each of the seven groups into natural and artificial looking movements.

In a control paradigm three consecutive test blocks were presented, followed by one training block and another test. This control experiment served to test whether spurious learning occurred during the presentation of the test trials.

Subjects
We tested 14 naive subjects in the discrimination and 7 naive subjects in the control paradigm. All subjects had normal or corrected-to-normal vision. They were tested individually.

Results

All subjects perceived every biological motion pattern as a human being, performing different kinds of exercises. Yet the interpretation of the underlying action was very different between subjects.

Figure 1b shows the averaged results over all 14 subjects. It can be seen that there is a gradual improvement in performance in the test blocks starting from 62% to about 85% in the third test. In contrast to the first test session, the correct responses in the first training block were already way above chance level ($> 80\%$) and there was no significant improvement in the second training block.

A one-way ANOVA for repeated-measurements comparing pretest, posttest 1 and posttest 2 showed that the observed increment in performance was significant ($F_{13} = 16.3; p < 0.001$). A Wilcoxon signed rank test was performed on the data comparing the results of the pretest with the first posttest. The results of this analysis revealed a significant difference between these two blocks ($p < 0.001; W+ = 1.5; W- = 76.5$). No significant difference was observed between

posttest 1 and posttest 2. However we found that the performance of the second posttest tended to be slightly higher compared to the first posttest.

The results for the test blocks sorted by the categorization judgments (natural/artificial) are presented in figure 2a. The mean percentage of correct responses separated for the groups of morphs that subjects classified as natural are shown in black, while the performances for the artificial looking movements are presented in gray. The gradual improvement during the test sessions seems to be uniform for movements categorized as natural looking as well as for movements the subjects classified as artificial. A two-way ANOVA for repeated-measures with the factors condition (Pre, Post1 and Post2) and category (natural / artificial) was performed on the data obtained from the categorization judgements. The results of this analysis show a significant main effect of condition ($F_{13} = 13.3; p < 0.001$) but no mean effect for category ($F_{13} = 0.9; p > 0.1$) and no significant interaction ($F_{13} = 0.7; p > 0.1$).

The results of the control paradigm are presented in figure 2b. It can clearly be seen that there is no improvement in performance over the three consecutive test blocks. The probability for correct responses was around 60% in all blocks. A one-way ANOVA comparing the three pretests revealed no significant difference ($F_6 = 0.7; p > 0.5$). However, if we compare the third pretest with the first posttest in a Wilcoxon test, we find a significant difference ($p < 0.05$).

Discussion

We have investigated the ability of human observers to learn the discrimination between new complex artificial stimuli. At first, we found that all subjects were able to learn to recognize completely new biological motion patterns.

Starting slightly above chance level, the subjects reached a level of about 85% correct responses after the two training sessions. Even one training block was sufficient to significantly increase the performance. However, if we tested subjects on three consecutive test trials without feedback, the number of correct responses did not increase. This suggests, that intermediate training together with feedback signaling whether the subjects answers were correct or not, was critical to improve performance. That is consistent with other studies, showing that discrimination performance can be improved by an "easy-to-hard" procedure. Subjects, that are first exposed to easy, highly separated discriminations along one dimension perform much better on subsequent more difficult discriminations along the same dimension. A possible explanation is that first presenting the easy discrimination allows humans to allocate attention to the relevant dimension [7].

With the help of the technique of motion morphing we were able to generate a pattern space in which the spatio-temporal similarity of the movements could be controlled by the weights of the prototypes. The fact that the performance for the first training block was much better than that for the first and second test block is consistent with the fact that the training patterns were more dissimilar to the center stimuli than the test patterns.

The main interest of this study was to investigate whether humans purely rely on stored innate templates to discriminate biological motion. The results of

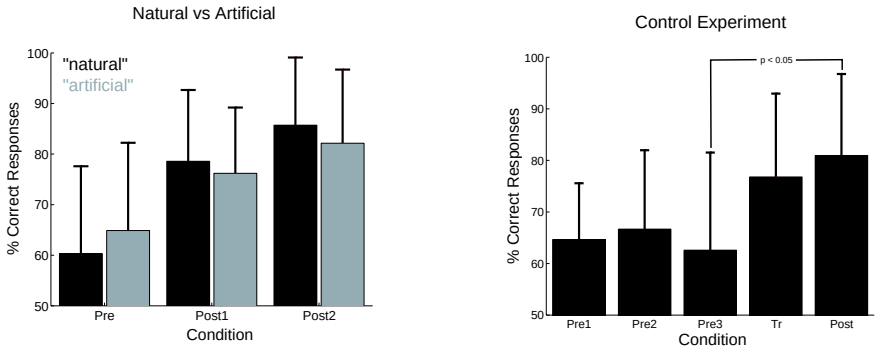


Figure 2: Results of experiment 1. Shown are the mean percentage of correct responses (+ s.e.m.) for the pretest and the two posttests. The data is split up in the performance for the groups, subjects classified as natural looking (black) and artificial looking (grey). (b) Results of the control experiment. Shown are the mean percentage of correct responses (+ s.e.m.) for the five different blocks.

our experiments seem to suggest that new templates can be learned. It seems unlikely that the recognition was purely based on innate templates, otherwise subjects should have more difficulties to discriminate between movements that appeared artificial than between those they classified as natural looking.

Our results clearly indicate that learning of complex biological motion patterns is possible and suggest that humans do not purely rely on stored innate templates. However, we cannot exactly specify the underlying learning process. For example, it remains unresolved whether the subjects exploited local or global features of the biological motion pattern to improve their performance. This motivates further studies that include more precise control of the strategies that subjects use in order to perform the learning task.

References

1. G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14:201–211, 1973.
2. L.T. Kozlowski and J.E. Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics*, 21:575–580, 1977.
3. R. Fox and C. McDaniel. The perception of biological motion by human infants. *Science*, 218:486–487, 1982.
4. B.I. Berenthal, D.R. Proffitt, and S.J. Kramer. Perception of biomechanical motions by infants: implementation of various processing constraints. *Journal of Experimental Psychology: Human Perception and Performance*, 13:577–585, 1987.
5. M.A. Giese and T. Poggio. Synthesis and recognition of biological motion patterns based on linear superposition of prototypical motion sequences. *Proceedings of the MVIEW 99 Symposium at CVPR, Fort Collins, CO*, 1999.
6. M.A. Giese and M. Lappe. Measurement of generalization fields for the recognition of biological motion. *Vision Researech*, 42:1847–1858, 2002.
7. M. Ahissar and S. Hochstein. Task difficulty and the specificity of perceptual learning. *Nature*, 387:400–404, 1997.

Tracking Human Hand Movements by Fusing Early Visual Cues

Axel Steinhage

Infineon Technologies AG, Corporate Research, Systems Technology, D-81730 Munich
Tel: +49-(0)89-234-55181, Email: Axel.Steinhage@infineon.com

Abstract. We describe an approach for the autonomous detection of skin-colored moving objects in man-machine-interaction scenarios. Based on low-resolution video images from an optically and mechanically uncalibrated low-quality camera, a simple image-processing algorithm extracts two visual cues from the scene: color and movement. By fusing these cues in real time, an implementation of the approach detects and tracks important scene-elements such as the moving hand or the face of a human interaction partner. The system builds a substantial part of an upcoming multimodal man-machine interaction system for mass-market applications.

1 Man-machine-interaction for mass-market applications

Currently, keyboard, mouse and text-based output on the monitor are still the most common means of communication between man and computer. However, regarding the tremendous development in speech recognition, image processing and virtual reality, it becomes obvious that in the future it will be possible to interact with machines by means of more natural communication channels such as speech, gestures and mimics. Although highly specialized solutions do already exist, the real breakthrough will happen once the systems become cheap, standardized and robust enough to be integrated in mass-market devices available to everyone (see Fig. 1). A major driver for this development will be the availability of robust recognition techniques which put only low demands on the hardware and processing power. These techniques should require only a minimum of calibration and adaptation by the user and no specific setup of the environment. In this paper we describe such a robust recognition technique which is able to track dynamic human hand movements based on ultra low-resolution video images from an optically and mechanically uncalibrated webcam. The aim is to implement a means of specifying and selecting objects displayed on a computer monitor just by pointing at them (see Fig. 1 and [2]).

2 An image-processing algorithm for extracting color and movement

The basic idea of our approach is the following: dynamic hand gestures are characterized by two basic features which correspond to so-called *early visual*

Fig. 1. Possible application of multimodal man-machine-communication in a *virtual shop* scenario: the user communicates with an artificial personal assistant on his tv-set (here in the lower right corner of the screen) by means of speech, gaze, pointing- and head-gestures using uncalibrated cheap devices like webcam and microphone. A combination with a video-conferencing system for the simultaneous communication with human partners (upper right) is planned.

cues in neurobiology (see also [1]): the object to track (i.e. the hand) has a specific color (i.e. the color of skin) and it moves in the image with a characteristic speed. Of course, there are many more high-level cues, such as the form, texture or specific trajectory which characterize a moving hand [3]. However, in order to be fast and robust, we stick with the simplest approach that still does the job. For similar reasons we want to get along with video from a monocular webcam only. As capture format we choose 80x60-RGB video images from a camera standing at an arbitrary position besides or on top of the monitor. The only requirement is that the hand is visible by the camera when it points to any position on the screen.

In the following, we describe our detection algorithm in detail. At first, the video stream is transformed from RGB to the Hue-Saturation-Intensity (HSV) color model in which H defines the so-called *color-angle* independently of the overall intensity (compare also with [4]). The full 360° color-angle is mapped onto the interval $H \in [0, 1]$. Human skin has an empiric color angle around 0°, so we generate a binary image $\mathbf{I}_{\text{skin}}(t)$ in which all pixels s in a small range $\cos(2\pi H_s) > c_{\text{skin}}$ around the skin-color angle are set to $I_{\text{skin}}(\boldsymbol{r}_s) = 1$ while all other pixels k vanish (i.e. $I_{\text{skin}}(\boldsymbol{r}_k) = 0$). The vector $\boldsymbol{r}_i$ denotes the position of the corresponding pixel i in the image. From the definition of the HSV-color model it follows that for low intensity and saturation, the color angle is not well defined. In the sunlight with high intensity V, white colored objects tend to have

a color angle similar to the human skin. Therefore, we ignore pixels that exceed an empirically chosen constant range for S and V. Summarizing, we have:

$$I_{\text{skin}}(\boldsymbol{r}_i) := \begin{cases} 1 \text{ if } (\cos(2\pi H_i) > c_{\text{skin}}) \text{ and } (S_i > c_{\text{S}}) \text{ and } (c_{\text{V1}} < V_i < c_{\text{V2}}) \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

Next, we extract movement from the intensity-image V using a dynamics

$$\dot{\mathbf{I}}_{\text{mean}}(t) = \alpha\,(\mathbf{V}(t) - \mathbf{I}_{\text{mean}}(t)) \quad (2)$$

which, depending on $\alpha > 0$, follows the temporal change of the intensity image $\mathbf{V}(t)$ over a time scale $\tau = \alpha^{-1}$. In the difference image $\mathbf{I}_{\text{diff}}(t) = |\mathbf{V}(t) - \mathbf{I}_{\text{mean}}(t)|$, only those areas appear that change on a time scale faster than τ. Hence, the image $\mathbf{I}_{\text{diff}}(t)$ represents movement and the static background is suppressed (see Fig. 2, upper right panel).

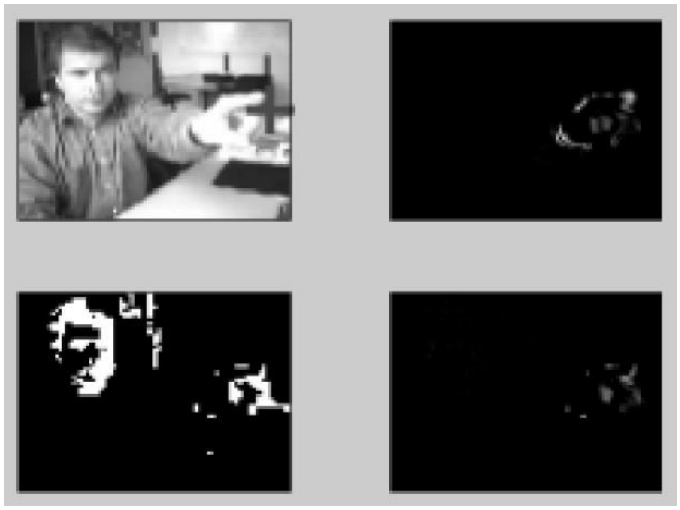


Fig. 2. Original video (upper left) in 80x60 resolution RGB24 captured by a webcam on top of the monitor. The skin-image $\mathbf{I}_{\text{skin}}$ (lower left) and the difference image $\mathbf{I}_{\text{diff}}$ (upper right) are multiplied to obtain the fused image $\mathbf{I}_{\text{fuse}}$ (lower right). The position of the maximum of the fused image is fed into the tracking dynamics and marked with a cross in the original image (upper left).

We are interested in movement of skin-colored objects only. Therefore, we fuse the color information with the movement cue. Based on the binary nature of the skin-color image $\mathbf{I}_{\text{skin}}(t)$, this can simply be done by multiplying the cue-images pixelwise:

$$\mathbf{I}_{\text{fuse}}(t) = \mathbf{I}_{\text{skin}}(t) * \mathbf{I}_{\text{diff}}(t). \quad (3)$$

In this fused image only those pixels appear that move and are skin-colored. We estimate the overall amount of movement in the fused image by calculating the number $S(t) = \sum_i (I_{\text{fuse}}(\boldsymbol{r}_i, t) > I_{\min})$ of pixels which exceed a level $I_{\min}$.

Finally, we simply track the maximum of the fused image by feeding its pixel-position $\boldsymbol{r}_{\max}(t) = \arg\max_i I_{\text{fuse}}(\boldsymbol{r}_i, t)$ into a fixed-point dynamics:

$$\dot{\boldsymbol{r}}_v(t) = \lambda \frac{\tanh(S(t) - S_{\min}) + 1}{2} (\boldsymbol{r}_{\max}(t) - \boldsymbol{r}_v(t)) \tag{4}$$

Herein, $S_{\min}$ is a parameter which guarantees that the movement is only tracked, if it exceeds a certain amount, i.e. if the number of active pixels in the fused image exceeds a given threshold. The dynamics (4) ensures that only continuous, smooth hand movements on the time scale λ^{-1} are tracked while spontaneous jumps (e.g. due to sensor noise) are filtered out.

3 Cursor control by means of pointing gestures

By setting the parameters appropriately (for their concrete values see next section), the algorithm can find and track any skin-colored moving object in the video image. However, for man-machine-interaction tasks, the machine must know the pointing direction in the current scene, i.e. in real world coordinates. Therefore, a method must be found which transforms the video-image-based coordinates $\boldsymbol{r}_v(t)$ into normalized coordinates $\boldsymbol{r}_s(t)$ on the computer screen.

An exact form of this transformation could be derived on the basis of the relative position of the camera with respect to the screen and the distance between hand and camera. While the former requires a mechanical calibration of the camera, the distance information can only be obtained by using complex algorithms like optical flow analysis or by means of a stereo camera system.

However, for the applications we have in mind, neither a stereo camera nor an exact mechanical calibration is possible: the system should work already after just placing a webcam on top of the monitor. On the other hand, for the virtual shop scenario no high precision cursor control is required as only objects which cover large portions of the screen are to be selected by the pointing gesture. Therefore, we implemented a very simple method similar to the classical mouse control: we assume that all possible coordinates $\boldsymbol{r}_v(t)$ cover roughly a rectangular region in the video image and map this region onto the rectangular screen by the following transformation:

$$\boldsymbol{r}_s(t) := \frac{\boldsymbol{r}_v(t) - \boldsymbol{r}_{\text{lo}}}{\boldsymbol{r}_{\text{hi}} - \boldsymbol{r}_{\text{lo}}} \tag{5}$$

The vectors $\boldsymbol{r}_{\text{lo}}$ and $\boldsymbol{r}_{\text{hi}}$ represent the lower left and the upper right corner of the rectangle in the video image. These vectors are determined during an initial calibration phase in which the user is asked to point to the lower left and upper right corner of the monitor. The normalized screen coordinates $\boldsymbol{r}_s(t)$ are used to control the position of the mouse pointer of the operating system.

4 Experimental results

We have implemented the algorithm in the form of a small MATLAB program (30 lines of code) using the VFM-capture plug in [5] and a Winnov PCMCIA-

Camera on Microsoft Windows. Using a video-resolution of 80x60 pixels and setting the parameter values $c_{\text{skin}} = 0.85, c_{\text{S}} = 0.05, c_{\text{V1}} = 0.3, c_{\text{V2}} = 0.9, \alpha = 0.9, I_{\min} = 0.1, \lambda = 0.1$ and $S_{\min} = 20$, hand movements can be tracked robustly independent of the user's tone of skin, the actual lighting conditions, the background and the position of the camera: as long as the hand is in the camera's field of view and represents the only skin-colored moving object, its movement is tracked (see Fig. 3). Due to the extreme simplicity of the algorithm, problems

Fig. 3. Robust hand-tracking in the presence of skin-colored but static distractors (upper left), with different camera position (lower left), for different persons (upper right) and with difficult lighting conditions (lower right).

arise when in addition to the hand another skin-colored object moves through the field of view. In that case, the system may switch from the hand to tracking the distractor. In practice this happens, for instance, when the user adjusts his seating position and moves his face or when other people move behind the user. However, as the user can observe the state of the tracker at any time, he can attract the "attention" of the system again by waving the hand anytime a tracking error is detected. In an interactive man-machine-communication scenario this behavior appears to be quite natural.

The transformation from video image coordinates to screen coordinates works relatively robust: in all experimental runs it was possible to use the screen coordinates $\boldsymbol{r}_s(t)$ for controlling the windows mouse cursor. By reducing the task to the selection of one of nine rectangular areas on the screen, even naive users, who were unfamiliar with the operation of the system, were immediately able to select the corresponding regions by means of pointing gestures (see Fig. 4).

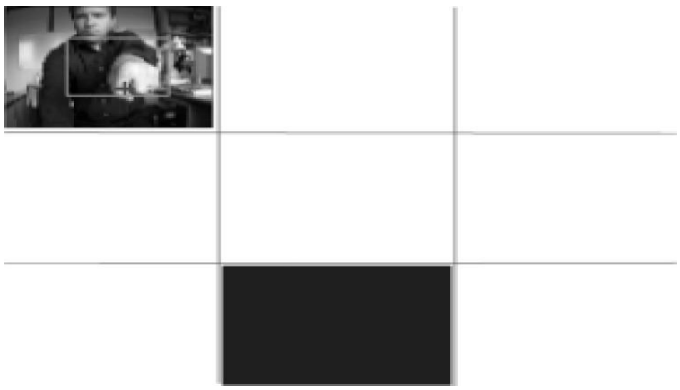


Fig. 4. Selecting rectangular areas on the screen by means of pointing gestures (here, the lower middle region was selected). The rectangle on the video image (upper left) indicates the region $(\boldsymbol{r}_{\mathrm{hi}}, \boldsymbol{r}_{\mathrm{lo}})$ resulting from the calibration phase.

5 Conclusion and outlook

We have presented a simple algorithm for tracking skin-colored moving objects based on the visual cues color and movement. The robustness against varying lighting conditions, camera position and background allows non-experts to use the algorithm in uncontrolled real world applications. The algorithm operates on low level sensor information (images) only and can be implemented in a highly parallel manner. The concept of cue-fusion by multiplication allows for a seamless integration of additional information such as a detector which separates head- from hand-movement.

Our work proves that high level behavior such as the robust visual tracking of specific objects can be generated by the direct fusion of low level (i.e. *early*) sensor information.

Future work will deal with the separation of head- from hand-gestures by means of additional cues and the integration of the algorithm into a multimodal man-machine-interaction scenario.

References

1. S.W. Lee, H.H. Bülthoff, T. Poggio (Eds.). Biologically Motivated Computer Vision, Proceedings of the first IEEE Int. Workshop BMCV 2000, Springer Verlag, 2000
2. VisionIC-Intelligente Vision-Plattform für den Massenmarkt, German Ministry for Science and Education BMBF, grant number 01 M 3127 B, 2002
3. J. Triesch and C. von der Malsburg. A system for person-independent hand posture recognition against complex backgrounds. IEEE Transactions on Pattern Recognition and Machine Intelligence, 23(12):1449-1453, Dec. 2001.
4. G. R. Bradski. Computer Vision Face Tracking For Use in a Perceptual User Interface. Intel Technology Journal Q2, 1998
5. F. Pezeshkpour. Vision for Matlab. http://www.sys.uea.ac.uk/~fuzz/projects.html.

Action and perception

Prediction of Rapidly Changing Environmental Dynamics for Real Time Behavior Adaptation using Visual Information

Emilia Barakova and Tino Lourens

GMD-Japan Research Laboratory
2-1 Hibikino, Wakamatsu-ku
Kitakyushu, 808-0135, Japan
http://www.gmd.gr.jp
{emilia.barakova, tino.lourens}@gmd.gr.jp

Abstract This paper features a method for acting in a real world environment with rapid dynamics, based on behaviors with different complexity, that emerge from: (1) direct sensing (2) on-line prediction about the future development of the environmental dynamics, and (3) internal restrictions derived by the robots strategy. The method is built upon the understanding of perception as dynamic integration of sensing, expectations, and behavioral goals, which is necessary when the environmental dynamics depends also on other intelligent agents. Two central aspects are considered: prediction of the development of environmental dynamics and the subsequent integration. The integration captures the dynamics of processes that happen in different temporal intervals but relate to the same perception. A real time object tracking method that is used is briefly described. Experiments made in a RoboCup environment with physical robots illustrate the plausibility of the method.

1 Introduction

In real-time dynamic environments reactive control has proven to be advantageous to the usage of symbolic representations and world modeling alone [1]. Reactive control relates direct sensing to robot actions, reflecting the understanding of perception as sensing. In this work is argued, that perception has to be considered as dynamic integration of factors related to the surrounding environment as sensing and predictions about the expected changes in the environmental dynamics, and the internal constraints of the robot as derived mainly by its behavioral goals. The internal constraints are also shaped by the robots physical body and the predefined constraints of the environment. Such an assumption is by far more realistic when the underlying environmental dynamics can deliberately be changed by other intelligent agents.

The integration is worked out within the application framework of the RoboCup scenario: competing robot teams in a soccer game. Since the goals of the two teams are obviously incompatible, the opposite team can be seen as a dynamic and obstructive environment [5]. The context of the environment narrows the

possible specter of plausible actions. Since sensor readings are available and behavioral goals are definable during the robots operation, the prediction of the opponent players behavior (which in a general context is a prediction of the dynamics of the environment) will further determine the best action.

Gross et al.[4] and the preceding works of Kosslyn [6], Moeller [8], and Pfeifer & Scheier [9] emphasize on the interrelated nature of action and perception, and lead towards an anticipatory model of perception, i.e. perception defined by the anticipatory action. Gross et al.[4] realize an internal anticipation and evaluation of several alternative sensory-motor sequences as a basis for an action-oriented perception. In addition, search methods are proposed, that will help selection of anticipative action. The performance of this neuro-biologically plausible model showed good results in navigation tasks in a static environment.

In dynamic environments however, the behavioral goals, which determine rather strategy-defined than logically straightforward actions, is a substantial part of the perception-action cycle. For instance, often in a real world the prediction of what is going to happen will determine one type of behavior, but goals, which have to be achieved can lead to completely different behavior.

To enable prediction, visual input data needs to be processed. The most widely used approach for object recognition in RoboCup is assigning pixels to color classes by thresholding, segmenting images by color, and assigning objects to so-called blobs in the segmented data [2]. It is clear that thresholding needs adjustment in different environments and that blob assignment to objects works only under highly restricted conditions. A newly developed approach that is robust under lighting conditions and that makes use of knowledge from the environment is proposed.

This paper is organized as follows: Section 2 outlines our hypothesis and approach. In Section 3 the vision method that is developing towards the state-of the art requirements of the sensing system is described. Section 4, features the prediction method. Some results with data from real games are shown. Section 5 outlines the integration process and finalizes the paper.

2 Perception-action model

The suggested perception-action scheme (Figure 1) accentuates on three elements of the action-oriented dynamic perception: direct Sensing, Predictions (expectations) and Behavioral goals and constraints. Sensing denotes the information, that is directly recorded by the sensors. It refers to the instant moment. Predictions are made on the basis of the on-line learned information about the environmental dynamics during the recent history and describe the expectations about its future development, i.e. it describes a future time event. The Behavioral goals (and constraints) reflect the available knowledge to the robot about the short-term goals it has to fulfill by taking the restrictions of the environment and its physical body into account.

Moreover, since perception is an intrinsically active process, it guides the actions of the robot and, conversely, the actions can take place in order to capture sensory information. The integration of the instant perceptions in the context

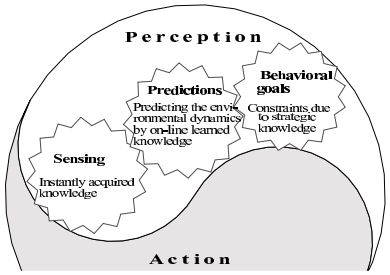


Figure 1. Proposed scheme of the perception-action interplay. The dynamic nature of this cycle is not explicit in this scheme, but has the following meaning: A history of sensing triggers predictions, the predictions shape the short-time behavioral goal which determine the action. Conversely, the action can be chosen to gather specific sensory information.

of the current situation (e.g. learned experiences and prediction about future development of the environmental dynamics) is the ultimate aim to be achieved.

The plausibility of this model has few aspects. First, it is closer to the actual nature of the perceptual process. Second, the multi-agent dynamic environment adds another degree of complexity to the behavior-oriented robotics: the robots action depends on various moving objects, some of which can commit deliberate changes into environmental dynamics due to opposite behavioral goals.

3 Object recognition

The proposed system for object recognition is build on a novel solution of real time object recognition. A combination of color and spatial reduction of image data insures a strong reduction of the visual information stream, and eases real time processing. Due to the fast dynamics of the environment this reduction is advantageous to all existing object recognition and tracking methods in RoboCup. The method insures real time perception and therefore makes prediction possible.

Detection of objects for a soccer playing robot comprises the following stages: color space reduction, spatial data reduction, color grouping, and object recognition.

Color space reduction transforms a full color image to an image of 7 different colors, white, black, green, blue, yellow, orange, and magenta. This reduction method is similar to the evaluation of colors by humans and therefore robust to lighting conditions, contrast differences, and moderate noise. Spatial reduction is performed to guarantee real time processing and is obtained by constructing a two dimensional perspective grid. The gridlines are obtained by calibration of both lens and camera angle with respect to the playground. The gridlines are separated by steps of 10 cm in the real world, which suffices, since all objects are at least 20 cm in size. The constructed grid is used to perform color segmentation. Objects, in turn, are evaluated as a set of segments and evaluated by their physical properties.

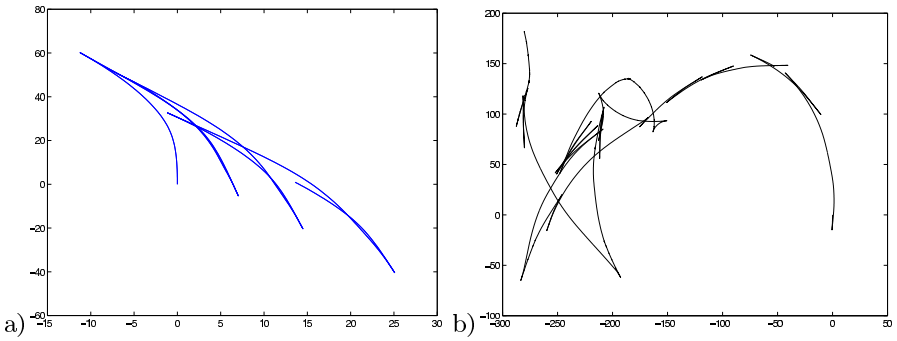


Figure 2. Robot path by no-attack a) and attacks from the left b).

4 Forecasting the environmental dynamics

Forecasting is build upon learning of the opponent robots behavior, or (initially) on the assumption of what the opponents behavior might be in the current scenario. Real-time prediction of the upcoming event based on on-line coming sensory cues is a very ambitious task. Instead, the forecast of a "type" or a "model" of the opponent behavior, is made during the ongoing game.

The on-line learned information of how the environmental dynamics tends to change, is a basis for predicting the complete upcoming event. The on-line coming sensory readings alone can be a base of a prediction of upcoming temporal history of the considered variables, if a drastic change does not take place. In environments with rapid dynamics unexpected changes are very possible. To cope with that fact, a representation on event level of abstraction is needed, together with incorporating the knowledge for the behavioral goals, the robot has to achieve.

More concretely, forecasting is build upon learning of the opponent robots behavior and the ball trajectory, or (initially) on the assumption of what the opponents behavior might be in the current scenario. For experimental testing an attacker-robot is used, that has to predict the goalie behavior. Within the RoboCup scenario, there are several possibilities for the behavior of a goalie: it opposes the movement of the ball; it opposes the movement of the ball and the attacker, it makes intermediate strategic movements (for instance in randomly chosen direction) in order to increase the complexity of the attackers decisions, the robot has unpredictable behavior (due to inaccuracy or malfunctioning).

It is important to say that the robot soccer programming and development environment provides processed sensory information in real time. For instance, instead of raw images, time series of distances and angles to the ball, goal, and recently other robots are available. In addition, the dynamics of behaviors like "following the ball" or "avoiding obstacles" can be recorded. A snapshot of sensor recordings and behavioral dynamics during a RoboCup game are shown in Figure 3a.

By combining the data from sensory and behavioral tracking the trajectory of the goalie is restored. The robot trajectories from many games and from simu-

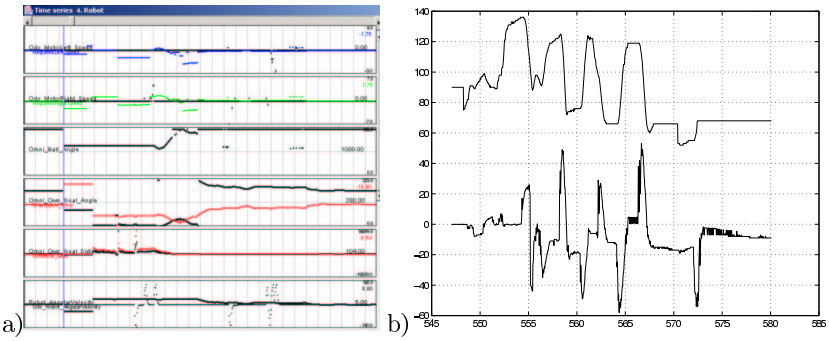


Figure 3. a) Recordings of sensor and behavior time series from a soccer game. b) Correlation between ball angle (top) and angular velocity of the robot (bottom).

lations are clustered by a neural gas algorithm [7] to represent various attacks or non-attack situations. In Figure 2 two typical trajectories of a goalie are shown: no attack (Figure 2a), and two subsequent attacks from left (Figure 2b). The trajectories, as clustered by the neural gas algorithm are used to derive the time the attacker has and the itinerary it has to take.

Once the dataset of trajectories are clustered, on-line classification of data takes place. The type of the attack situation is distinguished and the corresponding trajectory can be added to the attackers "normal" movement.

In the second experimental stage the relational trajectories are clustered, instead of trajectories that describe the movement of the robot. The relational trajectories describe the behavior of the goalie with respect to other moving objects: the ball and the attacking robot. Figure 3b illustrates that there is a strong correlation between the direction of movement of the ball and the response, i.e. the movement, of the goalie. Strong correlations can be found as well between the movement of the goalie with respect to the combined trajectories of ball and attacker.

5 Integrated perceptions in perspective

The dynamic interplay between the three elements (sensing, perception and behavioral goals) reveals the following stages: Initially, sensing and the straightforward behavioral goals are naturally integrated into the programmed behaviors. Acting in dynamic environments requires forecasting the tendency of environmental changes for adapting the behavioral outcome. Drastic changes in the robots surrounding indicate dynamics, caused by the actions of other intelligent agents or moving objects. They accentuate the need of incorporating the strategic knowledge, expressed as emerged short-term behavioral goals into the behavioral system. The three elements finally are expressed as trajectories or deviation from the trajectory, that will be taken by direct sensing only. Hence, the integration task transforms to combining the corresponding trajectories.

Previously, integration based upon temporal coherence principle has been proposed [3]. Due to its dependence of temporal cooccurrence of the information, the approach is not directly applicable for events that have happened in different temporal segments: sensing (current time), predictions (reflecting a future event, estimated on recent history), and (predefined) strategic knowledge. In this work the temporality is defined as relatedness to an event. After defining which part of every trajectory is related to the same event, the method proposed in [3] can be applied. In addition, the combination of the three trajectories has to cope with the problem of competing aims and is a self-contained problem to be solved in the future.

The suggested prediction method remains the central problem in this work. It has been put forward, that the prediction has to be made only within the interdependence of sensing and behavioral goals. The prediction captures the strategy of the goalie trough on-line analysis of the motion trajectories of the robot and its surrounding objects. To accomplish the on-line analysis, classification to previously learned types of behavioral models is made. The neural gas algorithm allows inclusion of a new class, if an unknown situation is encountered. As discussed before, this makes it suitable for on-line processing. An on-line version of the algorithm, that does not subdivide between clustering and classification will be considered. Additional experiments will be made that adapt to the extensions of the newly developed vision system.

References

1. Ronald C. Arkin and Tucker Balch. Cooperative multiagent robotic systems. In D. Kortenkamp, R. P.Bonasso, and R. Murphy, editors, *Articial Intelligence and Mobile Robots*. MIT Press, 1998.
2. E. I. Barakova and U. R. Zimmer. Dynamical situation and trajectory discrimination by means of clustering and accumulation of raw range measurements. In *Proc. of the Intl. Conf. on Advances in Intelligent Systems*, Canberra, Australia, 2000.
3. Emilia Barakova. An integration principle for multimodal sensor data based on temporal coherence of self-organized patterns. In J.Mira, R. Moreno-Diaz, and J. Cabestany, editors, *IWANN 2001*, LNCS, pages 55–63, part II, 2001.
4. H.-M. Gross, A. Heinze, T. Seiler, and V. Stephan. Generative character of perception: a neural architecture for sensorimotor anticipation. *Neural Networks*, 12:1101–1129, 1999.
5. H. Kitano, Y. Kuniyoshi, I. Noda, M. Asada, H. Matsubara, and E. Osawa. Robocup: A challenge problem for ai. *AI Magazine*, 18(1):73–85, 1997.
6. S. Kosslyn, N. Alpert, and W. Thompson. Visual mental imagery activates topographically organized visual cortex: Pet investigations. *Journal of Cognitive Neuroscience*, 5(3):263–287, 1993.
7. T. M. Martinetz, S. G. Berkovich, and K. J. Schulten. 'neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569, July 1993.
8. R. Moeller. Perception through anticipation-an approach to behavior-based perception. In *In Proc. New Trends in Cognitive Science*, pages 184–190, 1997.
9. R. Pfeifer and C. Scheier. From perception to action: the right direction? In *In Proc. PerAc '94*, pages 1–11, Las Almitos, 1994. IEEE Computer Society Press.

Effects of intracortical microstimulation in area MST on smooth pursuit

Uwe J. Ilg and Stefan Schumann

Kognitive Neurologie, Neurol. Universitätsklinik, Hoppe-Seyler-Str. 3, 72076 Tübingen
uwe.ilg@uni-tuebingen.de
stefan.schumann@uni-tuebingen.de

Abstract. The results of single-unit recordings from area MST during the execution of smooth pursuit eye movements suggest that these neurons code for the target movement within an external frame of reference. We support this assumption by the results of electrical stimulations within area MST. Stimulation affects the ongoing pursuit in a predictive and consistent manner according to the preferred direction at a given site. We observed more pronounced effects of stimulation if the target was absent during stimulation.

1 Introduction

One important consequence of the processing of dynamic visual scenes is the execution of smooth pursuit eye movements (SPEM). The spatial resolution of our visual system declines dramatically with eccentricity. Therefore we are constantly performing saccades (up to five per second) to direct the fovea towards items of interest in our visual surround and to utilize the high spatial resolution of foveal vision. Whenever such an item moves, we execute SPEM to maintain the retinal image of this item on or near the fovea. It is well established that the neuronal activity in the middle superior temporal area (MST) in the posterior parietal cortex of non-human primates is involved in the generation of SPEM (see Ilg 1997). Here, we address the question whether this area processes exclusively visual information or, alternatively, processes visual and extra-retinal information.

2 Pursuit-related activity recorded from area MST

Whenever neuronal responses are recorded during execution of SPEM, the origin of this activation has to be determined carefully. One possible source is self-induced retinal image motion if pursuit was performed across a visible background such as the borders of a computer monitor. To avoid this source, the pursuit experiments have to be executed in an absolutely dark laboratory equipped with a back-projection system onto a tangent screen for the visual stimuli. Another visual source is the retinal image motion of the target itself. Since SPEM can only be performed in the presence of a moving target, it is very difficult to avoid retinal image motion of the target itself. We

decided to use an imaginary target defined by peripheral visual cues. We initially trained the monkeys to track the centre of an hour glass. The imaginary target consisted of the hour glass with a blanked central area. We previously reported that rhesus monkeys could be trained to direct their eyes towards an imaginary target (Ilg and Thier 1999). Figure 1 shows the spike rates of an individual neuron in area MST during pursuit of the real and imaginary targets. A statistical analysis revealed that the responses to both targets were not significantly different. Based on the results from passive visual stimulation, we were able to exclude the possibility that the response to the imaginary target was due to stimulation of peripheral parts of the receptive field. It is important to note that we did not observe this independence of discharge rate from type of pursuit target when we recorded from the middle temporal area (MT). As others (Newsome et al. 1988; Thier and Erickson 1992), we conclude that individual neurons in area MST encode target movement in space based on a combination of retinal image motion and eye movement related signals.

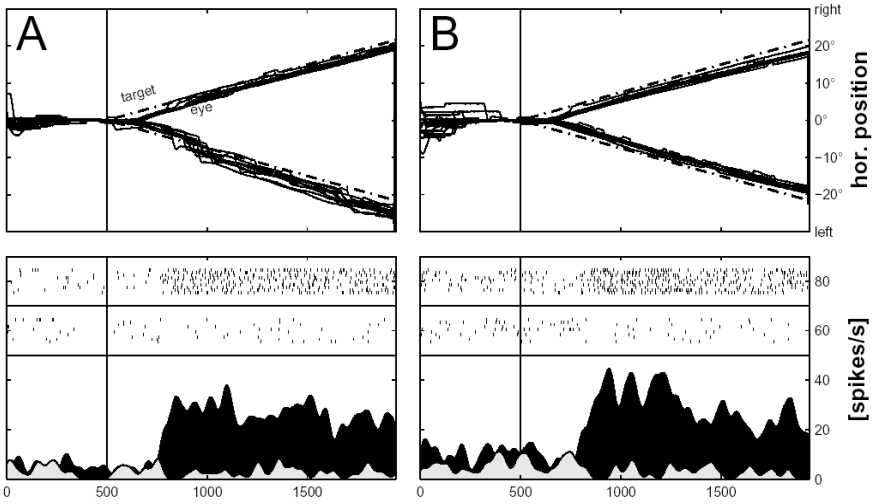


Fig. 1. Responses of a neuron recorded from Area MST during pursuit towards real (A) and imaginary target (B). The real target consisted of an hour glass (size 20°); in the case of the imaginary target, the central area (12°) was blanked. The upper row of raster display and the black spike density functions give the response during pursuit in the preferred direction; the lower row and the gray density functions the response in the non-preferred direction.

3 Intracortical microstimulation within area MST

To verify the above mentioned hypothesis, we applied intracortical microstimulation (ICMS) within area MST at sites with known preferred direction during the execution of SPEM. The specific location of area MST was determined during the above-mentioned single-unit study. We used ICMS in two different pursuit conditions. In

both conditions, a fixation target was presented in the centre of a tangent screen. Visual targets were back-projected onto this screen in absolute darkness; the borders of the projection screen were not visible for the monkey. After a fixation period of 1000 ms, the pursuit target started to move at a constant velocity (10 °/s). Stimulation current ranged from 40 µA to 120 µA. Stimulation started 200 ms after the onset of target movement and lasted for 200 ms. In the first condition, the moving pursuit target was visible while we applied ICMS. In the second condition, the target was switched off during the ICMS period. To determine exactly the stimulation effect, we sampled eye movement data during identical control trials lacking ICMS. Each condition was measured ten times; trials of all conditions were presented in randomized order. The initiation of SPEM is accomplished by the time of the initial saccade. Post-saccadic enhancement guarantees that eye velocity matches target velocity after the initial saccade (Lisberger 1998). Since this saccade always appeared during ICMS (mean saccadic latency 394 +- 96 ms, n=11840), we decided to use the post-saccadic eye-velocity (50 ms time window) to quantify the effect of ICMS. We compared eye velocity from stimulated trials with eye velocity from non-stimulated trials to calculate a vector of stimulation effect (VSE) for each measured pursuit direction. Figure 2 shows a typical example of the effects of ICMS in area MST during execution of SPEM. Single trials were aligned to the end of the initial saccade and median eye velocity traces were calculated. As Figure 2A shows, ICMS influenced post-saccadic eye velocity. If the direction of pursuit was in the preferred direction of the stimulated site, the stimulation yielded an increase in post-saccadic eye velocity. This effect of ICSM was stronger if stimulation was applied when the pursuit target was switched off, but did not change its direction.

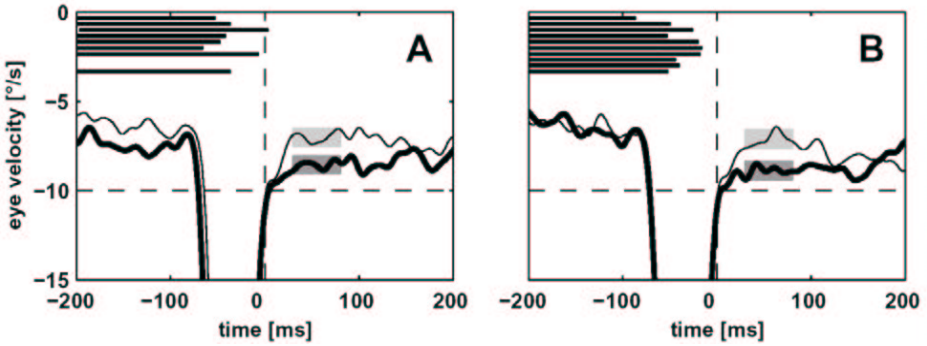


Fig. 2. Mean eye velocities during execution of SPEM in the preferred direction at the stimulation site. Single trials were aligned to the end of the initial saccade. Bold eye velocity profiles represent the eye velocity obtained from stimulated trials; normal profiles represent control trials without ICMS. The occurrence of ICMS in the individual trial is indicated by the horizontal lines. In A, the pursuit target was visible throughout the entire trial; in B the target was switched off during ICMS. The gray rectangles mark post-saccadic velocity.

Figure 3 shows the effects of ICMS of another site for pursuit in four different directions and the resulting VSE for each condition. The direction of the VSE was more or less independent of pursuit direction. By adding the four vectors, we

determined the mean effect of stimulation at this site. Note that the direction of the mean effect was very similar to the preferred direction of the stimulated site. We did not observe any effects of stimulation on the behavior of the monkey other than the reported modifications of eye movements.

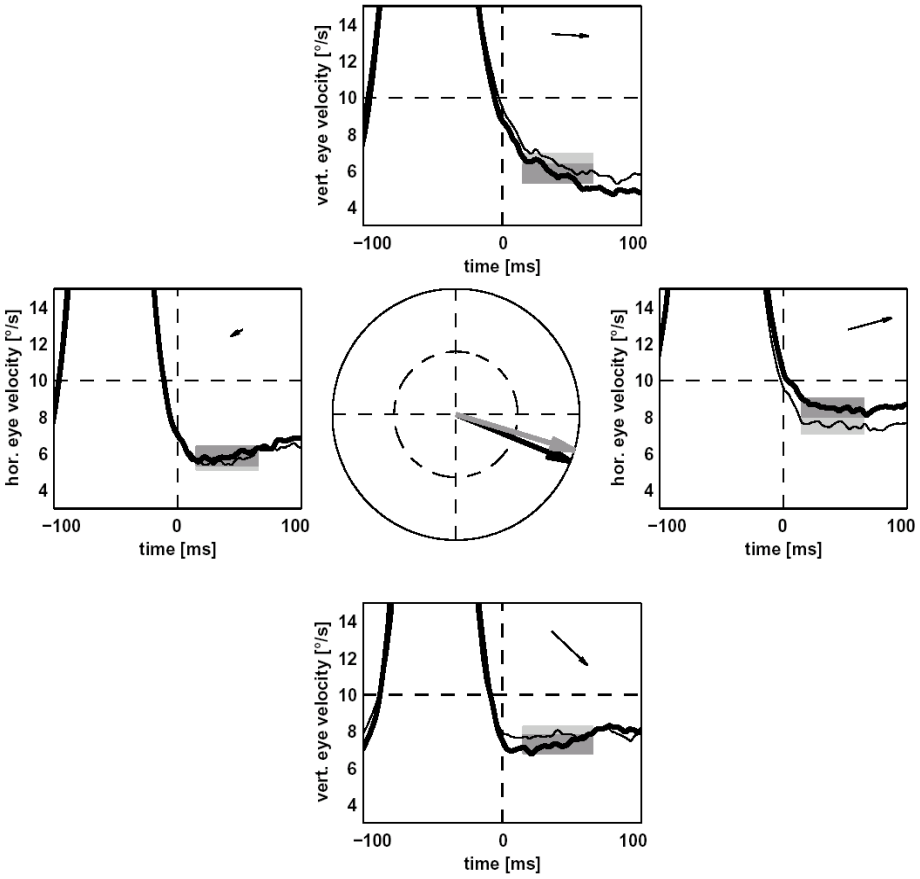


Fig. 3. Effect of ICMS for four different (left, up, right and down) pursuit directions. For details see Figure 2. Thin arrows show the direction and strength of the stimulation in the specific condition. The black arrow in the center represents the mean stimulation vector, the gray arrow gives the preferred direction at this stimulation site. Only eye velocity profiles obtained during stimulation in the absence of the pursuit target are shown.

4 Mean effects of ICMS

So far, we tested the effects of ICMS during SPEM in 74 stimulation sites in area MST of one rhesus monkey. Stimulation of 53 sites gave significant modulations in post-saccadic eye velocity. These sites were presumably all located within the

posterior bank and floor of the sulcus temporalis superior (STS) known as MST-l. We did not observe a significant difference in the preferred directions of these sites and the direction of the observed stimulation vector (t-test p = 0.068). The distribution of the angular difference between the two directions is shown in Figure 4.

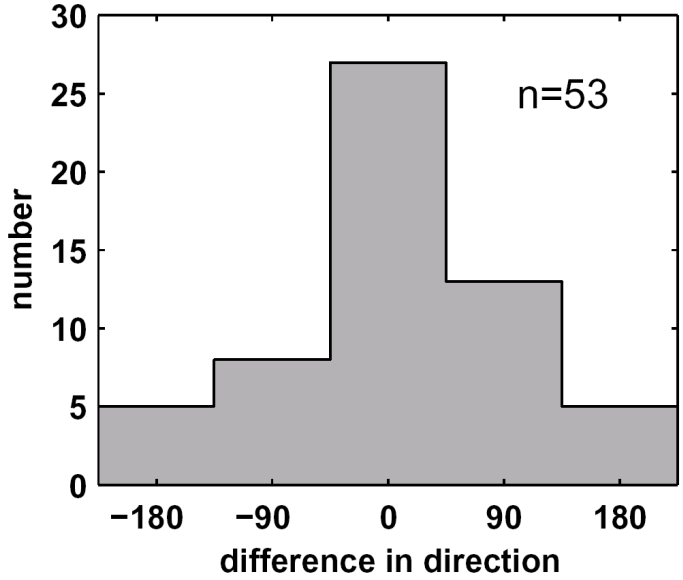


Fig. 4. Histogram of the angular difference between the preferred direction during execution of SPEM and the obtained mean stimulation vectors for 53 stimulation sites that had significant effects on the ongoing eye movements.

For all 53 stimulation sites, the mean absolute value of the VSE was 32% larger when the target was switched off during stimulation than when the target was visible for the complete trial. This difference in stimulation effect was significant (ttest, p=0.003).
In a previous study by Born and colleagues (2000), it was shown that area MT consisted of neurons with either wide-field or local field response properties. Stimulation of sites with wide field response properties resulted in a modulation in the opposite direction to the preferred direction of the stimulated site, whereas stimulation of sites with local motion characteristics resulted in a modulation in the preferred direction. For 42 out of the 53 sites, we determined the size tuning of the neuronal response observed at the given site to a moving stimulus during fixation. The vast majority of sites (37 out of 42) did not show an increase in the neuronal response with stimulus size. Conversely, these neurons gave a maximal response to a rather small stimulus, suggesting local motion characteristics. So our finding that the stimulation vector was in the same direction as the preferred direction of the individual stimulation parallels the earlier description for local motion sites (Born et al. 2000). The absence of wide field neurons might be for one of the following reasons: either our actual data sample is simply too small or area MST-l only contains local motion neurons. Nevertheless, in our present data sample, we did not observe the restriction

of the stimulation effect to an increase in eye velocity in ispisversive direction as reported by others (Komatsu and Wurtz 1989).

Conclusions

The conclusions of our results obtained from intracortical microstimulation in area MST are quite straight forward: smooth pursuit eye movements were accelerated in the preferred direction of the stimulation site. The effect was more pronounced if the pursuit target was invisible during stimulation. If the pursuit target was visible during ICMS, a combination of signals related to retinal image motion, eye movement, and artificial stimulation occurred. On the other hand, if the pursuit signal was switched off during ICMS, a combination of only eye movement related signals and artificial signals occurred. This observation further suggests the notion that the discharge rates of neurons in area MST represent target trajectory in space which is computed by a combination of retinal image motion of the target with eye and head movement related signals.

Acknowledgement

This work was supported by the Deutsche Forschungsgemeinschaft (SFB 550, A3 and Heisenberg Fellowship) and the Hermann and Lilly Schilling Foundation.

References

Born RT, Groh JM, Zhao R and Lukasewycz SJ (2000) Segregation of object and background motion in visual area MT: effects of microstimulation on eye movements. *Neuron* 26: 725-734

Ilg UJ (1997) Slow eye movements. *Prog Neurobiol* 53: 293-329

Ilg UJ and Thier P (1999) Eye movements of rhesus monkeys directed towards “imaginary” targets. *Vision Research* 39 (12): 2143-2150

Komatsu H and Wurtz RH (1989) Modulation of pursuit eye movements by stimulation of cortical areas MT and MST. *J Neurophysiol* 61: 31-47

Lisberger SG (1998) Postsaccadic enhancement of initiation of smooth pursuit eye movements in monkeys. *J Neurophysiol* 79: 1918-1930

Newsome WT, Wurtz RH and Komatsu H (1988) Relation of cortical areas MT and MST to pursuit eye movements. II. Differentiation of retinal from extraretinal inputs. *J Neurophysiol* 60 (2): 604-620

Thier P and Erickson RG (1992) Responses of visual-tracking neurons from cortical area MST-l to visual, eye and head motion. *Europ J Neurosci* 4: 539-553

Neuronal requirements for execution of smooth pursuit and motion perception

Jan Churan and Uwe J. Ilg

GRP der Universität München, Arzbacher Str. 12, 83646 Bad Tölz
churan@grp.hwz.uni-muenchen.de
Kognitive Neurologie, Neurol. Universitätsklinik, Hoppe-Seyler-Str. 3, 72076 Tübingen
uwe.ilg@uni-tuebingen.de

Abstract On the basis of two different motion stimuli, we were able to demonstrate that rhesus monkeys perceive these stimuli and are able to track these stimuli with their eyes. However, we did not observe directionally selective activation of neurons in area MT and MST, commonly believed to be important for visual motion processing. We conclude that both perception of motion as well as final sensorimotor processing are achieved in cortical areas beyond areas MT and MST.

1 Introduction

It is well established that the processing of visual motion in the middle temporal (MT) and middle superior temporal (MST) areas in the posterior parietal cortex of monkeys is closely related to the execution of smooth pursuit eye movements (SPEM) (for review see Ilg 1997 and chapter of Ilg and Schumann) as well as the perception of visual motion (e.g. Celebrini and Newsome 1994, 1995). The properties of individual neurons in these areas are very well suited for these tasks since they code for the direction as well as for the speed of a moving stimulus.
It was suggested that perception and action might depend on separate visual mechanisms (Goodale and Milner 1992). In order to ask whether this dichotomy also holds true for the processing of motion underlying motion perception and generation of smooth pursuit eye movements, we combined psychophysical, eye movement and single-unit response studies in awake and behaving rhesus monkeys. Specifically, we investigated whether directionally selective single-unit activity in areas MT and MST indicating the direction of a moving object is a necessary condition for the execution of SPEM as well as for the perception of motion. In addition to a first-order (fourier) motion stimulus, we used two other types of motion stimuli: a paradoxical second-order motion stimulus and a visual-auditory multimodal motion stimulus.

2 Types of motion stimuli

The paradoxical second-order motion type used in our experiments was the theta motion as described by Zanker (1993). In this stimulus, a rectangular patch of random dots moves over a dynamically flickering random dot background; the dots within the rectangle move in the opposite direction as the rectangle itself. Therefore, the local motion (motion of the dots) and the global motion (motion of the object) components are moving in opposite directions. As a control, we used the first-order motion stimulus, which consisted of an identical rectangle of coherently moving dots. It is important to note that the raw motion signal (number and velocity of the moving dots) is identical in these two stimuli. Only the relationship between the direction of dot movement and object movement differs between the two stimuli.
The visual-auditory multimodal motion was produced by a horizontal array of 48 LED and loudspeaker elements (distance between two elements: 0.95°). To generate the percept of motion, we activated the elements sequentially for 25 ms with a temporal gap of 25 ms. We activated either only the LED elements (visual motion), only the loudspeakers (white noise, auditory motion) or both (multimodal motion). This presentation resulted in an apparent motion of the stimulus at the velocity of 18,7 °/s.

3 Motion perception of rhesus monkeys

We trained three rhesus monkeys to a direction discrimination task of the various motion stimuli used here. The monkeys had to fixate a central fixation point. After a randomized time period, the motion stimulus was displayed while the monkeys had to maintain fixation. Following an additional delay, the monkeys had to report the perceived direction of motion by a saccade directed towards one of the two simultaneously presented saccade targets.
All monkeys learned to report correctly the direction (leftward vs. rightward) of the presented stimuli (85% correct responses for the first-order stimulus, 78% for the theta stimulus, 93% correct responses for the visual and visual-acoustic stimuli, 76% for the acoustic stimulus).

4 Smooth pursuit eye movements

Having shown that the monkeys were able to discriminate the motion direction of the theta stimulus, we asked whether monkeys are able to perform SPEM to the theta stimulus. This was previously demonstrated for human subjects (Butzer et al. 1997, Lindner & Ilg 2000). After fixation of a central stationary target for a random period of time, the monkeys had to track as precisely as possible the ramp-like movement of the stimulus (10°/s). After a brief period of training, the monkey performed SPEM to the first-order and theta-motion stimuli. However, the steady-state eye velocity gain in the case of theta-motion (average of 0.6 for 46 periods of measurement including

about 7000 trials) was significantly lower ($p<0.001$) than the gain obtained for SPEM to a first-order stimulus (average of 0.9 for 3 periods of measurement including about 1000 trials) (see Fig. 1). Periods of slow eye movements were interrupted by catch-up saccades which compensate for the insufficient eye velocity during the SPEM periods.

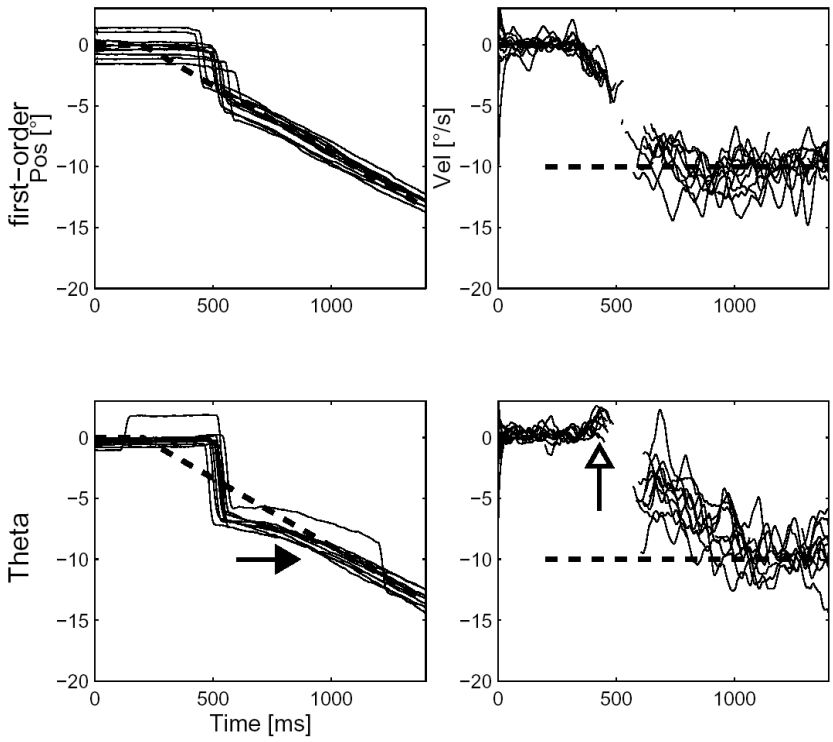


Fig. 1. Horizontal eye position and de-saccaded eye velocity elicited by a first-order (upper row) and by a theta-motion stimulus (lower row) moving at a speed of 10°/s. The monkey was able to perform steady-state SPEM (black arrow) to the two stimuli, however the SPEM was less precise when the theta stimulus was presented. Note that during initiation of SPEM (open arrow), the eye movements were transiently in opposite direction in case of the theta stimulus.

As indicated in Figure 1, the initiation of smooth pursuit eye movements directed towards a theta stimulus followed the movement of the individual dots, i.e. were opposite to the direction of the moving object. In a study of human pursuit, we quantified exactly the eye acceleration. Although the raw motion signal in fourier and theta motion was identical, the elicited acceleration was significantly smaller in the case of theta motion (Lindner and Ilg 2000).

5 Single-unit responses recorded from areas MT and MST

We recorded 38 neurons from area MT and 68 neurons from area MST in both monkeys performing the direction discrimination task. We only included those neurons which gave a direction selective response to the first-order motion stimulus. The receptive fields of the neurons recorded from area MT were slightly but significantly smaller than the fields from area MST. Besides this difference, we did not observe any other significant differences in the response properties of neurons from the two areas.

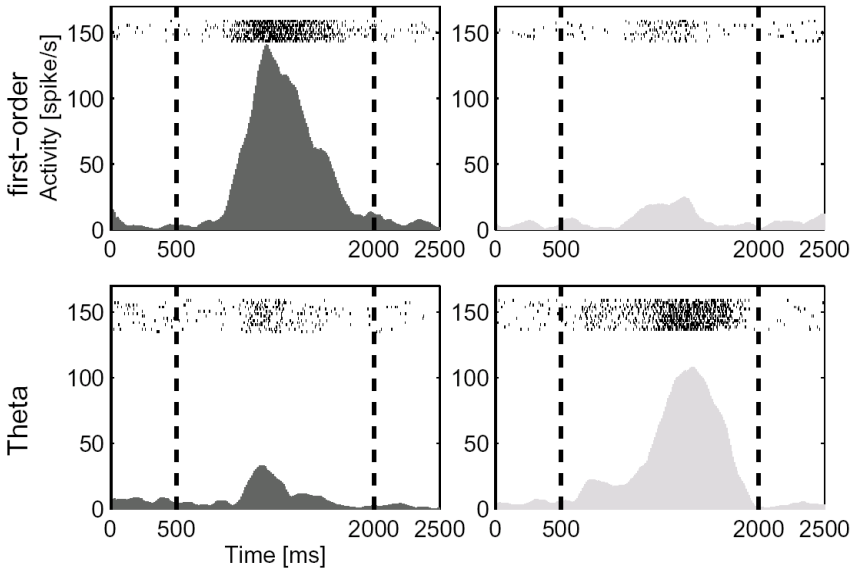


Fig. 2. Responses of a neuron from area MST to first-order and theta-motion stimuli shown as raster display and spike density function. The left column shows the responses elicited by leftward stimulus movement, the right column shows the responses elicited by rightward movement. The preferred direction of the neuron is apparently inverted for the theta-stimulus in comparison to the first-order stimulus. Despite this apparent inversion of preferred direction, the monkey reported correctly the direction of the moving stimuli.

Figure 2 shows the response of a typical neuron recorded from area MST. In the case of the first-order motion stimulus, the neuron showed a massive response to leftward motion. During presentation of the theta stimulus, the neuron responded to rightward motion. We made this observation in all 106 neurons recorded from areas MT and MST. Obviously, the neuron responded to the movement of individual dots in the display, not to the movement of the entire object. However, when we analyzed the sharpness of the directional tuning of the responses to fourier and theta stimuli, we found that the sharpness of the response to theta motion was reduced compared to fourier motion. This finding parallels exactly our finding related to the pursuit initiation elicited by the fourier and the theta stimuli.

It is important to note that area MT and area MST are organized in a retinotopic fashion. The motion information related to the object motion of the theta stimulus is possibly provided by correlated activation of neighboring parts of these areas which can be read out by a subsequent processing stage.
The other type of motion stimuli consisted of the apparent motion of visual, auditory and multimodal stimuli. As mentioned above, our monkeys were able to report correctly the direction of the stimuli. However, when we recorded the neuronal responses during this task, we only observed responses to the visual and visual-auditory moving stimulus as shown in Figure 3.

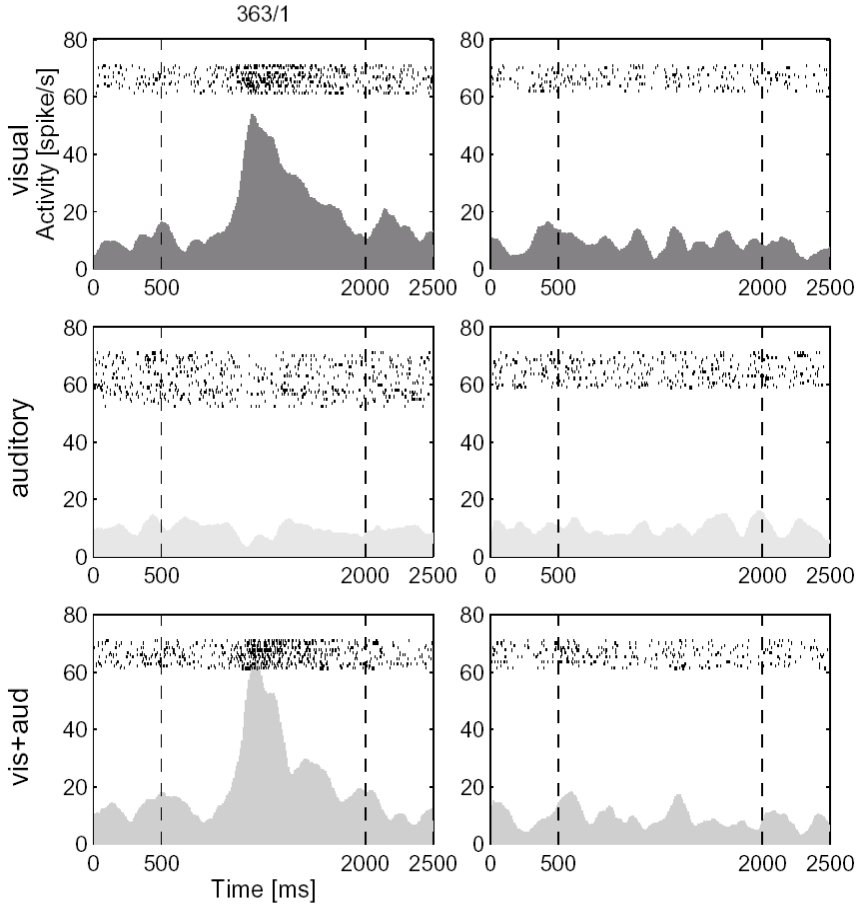


Fig. 3. Response of a typical neuron recorded from area MST to visual, auditory and visual/auditory motion shown as raster display and spike density function. Although there was no response to the auditory stimulus, the monkey reported correctly the direction of all moving stimuli.

Thirty-two of the 96 neurons examined in the multimodal motion task gave significantly direction-selective responses to the apparent motion of the visual and the

visual-auditory stimuli. We did not find a single neuron that gave a response to the auditory movement. Despite this lack of neuronal activity, the monkeys reported correctly the direction of the moving auditory stimulus.

Conclusions

The smooth pursuit eye movements of human subjects and rhesus monkeys follow in their steady-state phase the direction of the perceived object motion. This indicates that the visual motion processing underlying perception and sensorimotor integration depends on a common mechanism. Furthermore, the similarity in initiation of smooth pursuit eye movements and neuronal responses recorded from areas MT and MST suggest that these areas are part of this mechanism. However, our results show that rhesus monkeys can perform steady-state SPEM as well as motion perception tasks in the absence of explicit coding of object motion in the activity of neurons recorded from areas MT and MST. So we conclude that the perception of a moving stimulus as well as the generation of smooth pursuit eye movements reflects the achievement of a motion area located higher than area MT and MST in the hierarchy of cortical visual information processing.

Acknowledgement

The work was supported by the Deutsche Forschungsgemeinschaft (SFB 550, A3 and Heisenberg Fellowship) and the Hermann and Lilly Schilling Foundation.

References

Butzer F, Ilg UJ, Zanker JM (1997) Smooth-pursuit eye movements elicited by first-order and second-order motion. *Exp Brain Res* 115:61-70.

Celebrini S and Newsome WT (1994) Neuronal and psychophysical sensitivity to motion signals in extrastriate area MST of the macaque monkey. *J Neurosci* 14 (7): 4109-4124

Celebrini S and Newsome WT (1995) Microstimulation of extrastriate area MST influences performance on a direction discrimination task. *J Neurophysiol* 73: 437-448

Goodale MA and Milner DA (1992) Separate visual pathways for perception and action. *Trends in Neurosci* 15 (1): 20-25

Ilg UJ (1997) Slow eye movements. *Prog Neurobiol* 53:293-329.

Lindner A, Ilg UJ (2000) Initiation of smooth-pursuit eye movements to first-order and second-order motion stimuli. *Exp Brain Res* 133:450-456.

Zanker JM (1993) Theta motion: a paradoxical stimulus to explore higher order motion extraction. *Vision Res* 33:553-569.

Analysing Adaptive Behaviour from a Macroscopic Perspective[1]

Michel van Dartel, Eric Postma, and Jaap van den Herik

IKAT, Universiteit Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands
{mf.vandartel, postma, herik}@cs.unimaas.nl

Abstract. This paper investigates whether a macroscopic analysis enables the identification of universal properties of adaptive behaviour in situated agent (robot) models. In contrast to microscopic analysis, macroscopic analysis focuses on averaged properties of systems. For our purpose, a macroscopic analysis of adaptive systems is performed. The adaptive systems studied are evolutionary optimised foraging agents. The analysis reveals that the step lengths of the most successful agents are distributed according to a Lévi-flight distribution. Such a distribution constitutes a universal property of foraging behaviour that is encountered in many natural species. Hence, in this domain macroscopic analysis clearly facilitates the discovery of universal properties of adaptive behaviour. Generalising this conclusion, we believe that macroscopic analysis is complementary to microscopic analysis in the study of adaptive behaviour.

1 Introduction

In-depth analysis of simple agent models reveals many new insights into the processes underlying adaptive behaviour and situated cognition [see, e.g., 2, 3, 4, 8]. So far, analysis is only done at the microscopic level, in which the focus is on the successful behaviour of single agents only. Although microscopic analysis can lead to explanatory insights and testable predictions at an individual level, due to this specificity, generalisation of results is difficult. In contrast, macroscopic analysis is more suitable for identifying universal properties, i.e., properties characteristic of a class of systems. Macroscopic analysis ignores individual differences by averaging over large quantities of data. The application of macroscopic analysis in statistical physics led to successful extraction of universal properties of, for instance, DNA sequences, heartbeat rates, and weather variations [6,10]. A recent example of macroscopic analysis of natural behaviour is the study by Beekman *et al.* [1], who analysed foraging behaviour of Pharaoh ants. They revealed collective foraging behaviour to exhibit a phase transition from disordered to ordered foraging when the size of the colony was increased.

The research question addressed in this paper reads: Can macroscopic analysis extract universal properties of adaptive behaviour from situated agent (robot) models? To answer this research question we optimise the foraging behaviour of neural-network controlled agents using evolutionary-computation techniques. Next, we perform a macroscopic analysis on the foraging behaviour of the optimised agents.

---

[1] An earlier version of this paper was published in the Proceedings of the Fourteenth Belgium-Netherlands Artificial Intelligence Conference (BNAIC) 2002.

The outline of the remainder of the paper is as follows. In section 2, the foraging experiment is outlined. Section 3 presents the results of the macroscopic analysis. In section 4, the results of the analysis are discussed and related to other findings. Finally, the conclusion given in section 5 reads that macroscopic analysis facilitates the identification of universal properties of adaptive behaviour in agent models.

2 The foraging experiment

The foraging experiment is outlined in terms of the environment (section 2.1), the agent (section 2.2), and the evolutionary-computation algorithm (section 2.3).

2.1 The environment

The environment is defined as a $L \times L$ square with periodic boundary conditions (i.e., the environment is defined on a torus) containing n food elements. Randomly distributed dots over the environment represent the food elements. An agent collects food by walking over the food elements. Whenever a food element is collected, it is removed from the environment and replaced by a new one at a random location. In this way, the number of food elements remains constant throughout the experiment. Figure 1 is an illustration of the environment with randomly distributed food elements (dots) and the agent (circle).

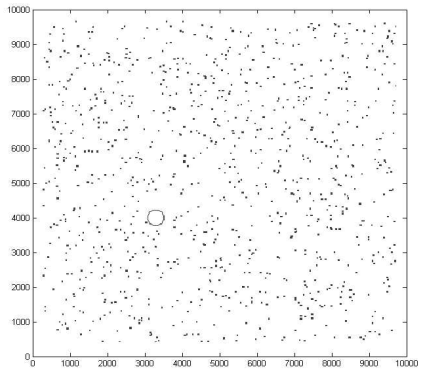


Figure 1. Illustration of the experimental environment consisting of an agent (circle) and randomly distributed food elements (dots). The values on the x- and y-axes are spatial coordinates ($0 \leq x,y < L = 10000$).

2.2 The agent

The agent performing the foraging task is controlled by a neural network and is defined in terms of sensor and brain.

Sensor. The sensor of the agent detects the nearest food element within its circular field of view with radius r. The sensor processes the nearest food element within the field of view only and is orientation sensitive. Defining the orientation of the agent by α and the

orientation of the nearest food element by β, the sensor activation I (i.e., the input) is given by the normalized one-dimensional Von Mises basis function [7].

$$I = \frac{e^{k\cos(\alpha-\beta)}}{e^{k}} + G(0.0, sd), \tag{1}$$

where k is a positive constant that is proportional to the width of the basis function. The Von Mises basis function is the spherical analogue of the Gaussian basis function. The normalisation constant e^k ensures that the maximal value of the first right-hand side term equals 1 when $\alpha = \beta$. The second term is a Gaussian-noise term (zero mean, standard deviation *sd*), modelling the intrinsic noise of neural systems. A food element is collected when the distance between the food element and the agent equals $0.1r$.

Brain. The brain (or controller) of the agent is a recurrent neural network with a single input I, H hidden nodes, and two output nodes. The input is connected to the hidden and to the output nodes. The hidden nodes have recurrent adaptive connections. Each connection can be switched on or off during the evolutionary process, while retaining its weight value (cf. [9]). Initially, all weights are assigned random values symmetrically distributed around zero on the interval $[-rw,rw]$, with $rw > 0$. The transfer function for the hidden nodes is the sigmoid *tanh* function that maps onto the interval $\langle -1,1\rangle$. The two output nodes control the agents' relative orientation and step size, respectively. The transfer functions for the output nodes are defined as follows. The output of the orientation node is multiplied by π. A modulo operation restricts the orientation to the interval $\langle -\pi,\pi\rangle$. The transfer function of the step-size output node is a semi-linear function $l = f(u)$ that maps negative values to zero and positive values u to the interval $\langle 0, uL/2\rangle$, with L the width/heigth of the square environment.

2.3 Evolutionary-computation algorithm

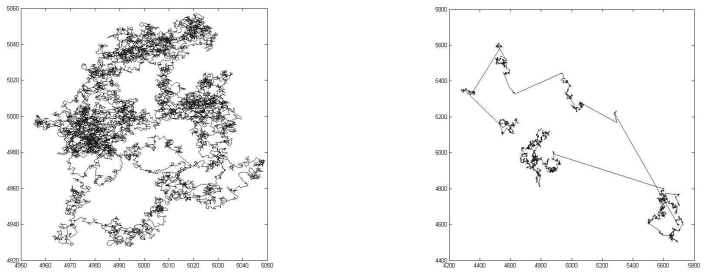
The weights of the neural network controlling the agent are optimised for foraging efficiency using a standard evolutionary-computation algorithm. The fitness function F is defined as follows.

$$F = \frac{1}{T}\left(\sum_{t=1}^{T} c(t) - \lambda \sum_{t=1}^{T} l(t)\right), \tag{2}$$

where t is an index for individual simulation steps ($t \in \{1,2,\ldots T\}$) with T denoting the total number of steps, $c(t)$ is a function that returns *1* if a food element is collected at step t and *0* otherwise, $l(t) = f(u,t)$ is the step length of the agent at step t, and λ is a positive parameter. The first term between the brackets favours food collection. The second term punishes long steps. The balance between the two terms is set by λ. All simulations are based on an evolution of 100 generations with a population size of 1000 agents. Evolution occurs using standard evolutionary optimisation techniques (see [9]).

3 Experimental results and analysis

A large series of experiments was performed to optimise the foraging behaviour of the agent. The simulations yielded various types of behaviour. Figure 2a and b show two typical examples of behaviours associated with high fitness values. The figures show the paths traced by the optimised agent. Although most optimised agents perform the random-walk behaviour shown in figure 2a, some agents exhibit the qualitatively different behaviour shown in figure 2b. A characteristic feature of these agents is that their local random-walk behaviour is occasionally interrupted by large jumps. As a result the area covered by these agents is much larger than the area covered by random-walk agents. The sudden jumps are known as Lévi flights [5,11,12]; they will be discussed below. Foragers adopting a Lévy-flight strategy outperform the agents using a random-walk strategy. Apparently, the Lévy flights are more effective in terms of foraging efficiency than the random walks.



a b

Figure 2. Illustration of (a) random-walk foraging behaviour and (b) Lévy-flight foraging behaviour. Both paths consist of 10.000 steps. It should be noted that the area covered in figure (b) is much larger than the area covered in figure (a). The values on the x- and y-axes are spatial coordinates ($0 \leq x,y < L = 10000$).

Our macroscopic analysis focuses on the quantification of the difference between random walks and Lévy flights in terms of a single parameter μ. The parameter is extracted from the probability density function (pdf) from which the lengths of the steps taken during foraging are drawn [11]. Concentrating on the probability of large step lengths, the tail of the pdf scales according to (cf. [12]):

$$P(l) = \frac{l^{-\mu}}{Z}, \tag{3}$$

with $P(l)$ representing the probability of a step of length l, and Z a normalising constant. The parameter μ is proportional to the rate of decay of the pdf with length l. For a Gaussian pdf that generates random-walk behaviour, the parameter μ is larger than 3.0. Lévy-flight behaviour is associated with $1.0 < \mu \leq 3.0$. These values of μ yield 'fatter' tails, leading to infinite variance and an undefined average of the pdf. In our agent, the pdf is generated from the step lengths produced by the output node. To perform our macroscopic analysis we created step-length histograms by running series of foraging simulations using optimised agents of both the random-walk and Lévy-flight

types. Figure 3a shows an example of a histogram so obtained. Subsequently, we analysed the (smoothed) tails of the histograms by fitting a linear regression line through the data points. The slope of the line is an estimate of the value of μ that underlies the behaviour of the two types of agents. Figure 3b displays the regression line for an agent that exhibits the Lévy-flight behaviour shown in figure 2b. The slope of the regression line is approximately equal to –2 (i.e., μ = 2). In terms of equation (3) this corresponds to $P(l) = l^{-2}/Z$.

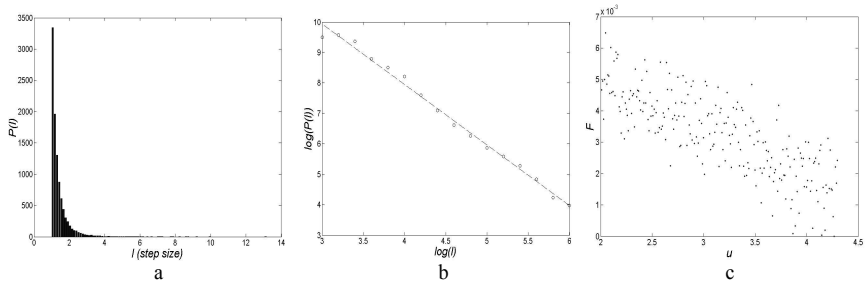


Figure 3. (a) Histogram of step lengths. (b) Log-log plot of the tail (*l >> 0)* of the histogram. The slope of the regression line is ≈ -2. (c) The fitness *F* as a function of the step-size distribution parameter μ for the 281 fittest foragers.

We performed a series of experiments with the parameter values: H = 2, r = 1, L = 10000, number of food elements = 100, rw = 0.5, k = 20, sd = 0.5, T = 10000, and λ = 0.00001. The experiments yielded a population of 1000 optimised foragers. Of these foragers, the 281 fittest ones performed a range of random-walk and Lévy-flight behaviours. The remaining 719 foragers employed various sub-optimal strategies such as foraging along straight lines. For each of the 281 fittest foragers a histogram (such as shown in figure 3(a)) was created from several runs of T steps each. A log-log plot of the tail of the histogram is shown in figure 3(b). Subsequently, the value of μ was determined for each histogram. The values of μ ranged from μ ≈ 3.5 to μ ≈ 2.0. Figure 3(c) plots the fitness of the 281 fittest foragers as a function of μ. Interestingly, the fittest foragers are associated with values closer to μ ≈ 2.0. Evidently, optimal fitness values are found near μ = 2, which is associated with Lévy-flight foraging behaviour.

4 Discussion

The macroscopic analysis of our model of foraging behaviour led to the extraction of a universal property of efficient foraging, i.e., Lévi flights as characterised by the universal exponent μ. A range of animals exhibits efficient foraging behaviour that is characterised by Lévi flights with μ ≈ 2: albatrosses, foraging bumblebees, deer, and amoebas [11, 12].

Microscopic analyses of agent models of adaptive behaviour explain behaviour on an individual level. For instance, such an analysis can reveal the dynamical processes underlying catching and avoiding behaviour in individual agents (see [4]). Generalisation to other types of agents and situations is difficult because of the idiosyncrasy of the agent-environment interaction. Since macroscopic analysis averages over many interactions, it obscures the details of the interaction, but uncovers generic

properties. The value of the exponent μ cannot be determined from the study of a single agent, but instead requires the averaging over many interactions and environments. However, once the macroscopic analysis revealed the value of μ and related it to Lévy flights, the characteristic foraging behaviour is readily recognised in the microscopic behaviour of the individual agent (see, e.g., figure 2b). Macroscopic analysis is therefore complementary to microscopic analysis. Agent models of adaptive behaviour have to be analysed at both levels to gain a complete understanding of the behaviour.

5 Conclusion

Using macroscopic analysis we extracted a universal property of foraging behaviour in artificially evolved agents, i.e., Lévy flights as characterised by $\mu = 2$. By doing so, we have shown that macroscopic analyses of agent models can identify universal properties of adaptive behaviour. Given this finding, and the successes of macroscopic analyses in statistical physics and other disciplines, we expect macroscopic analyses to generate novel insights into the universal properties of adaptive behaviour in artificial and natural agents.

References

1. Beekman, M., Sumpter, D.J.T., Ratnieks, F.L.W.,: Phase transitions between disordered and ordered foraging in Pharaoh's ants. Proceedings of the National Academy of Sciences, 98. (2001) 9703-9706
2. Beer, R.D.: Computational and Dynamical Languages for Autonomous Agents. In: Port, R., van Gelder,T. J. (eds.): Mind as Motion: Dynamics, Behavior, and Cognition. MIT Press, Cambridge, MA (1995)
3. Beer, R.D.: Dynamical approaches to cognitive science. Trends in Cognitive Sciences 4(3). (2000) 91-99
4. Beer, R.D.: The Dynamics of Active Categorical Perception in an Evolved Model Agent. Behavioural and Brain Sciences. Submitted (2001)
5. Gutowski, M. Lévy flights as an underlying mechanism for global optimization algorithms. Math-ph/0106003 (2001)
6. Havlin, S., Buldyrev, S.V., Bunde, A., Goldberger, A.L., Ivanov, P., Peng, C.-K., Stanley H.E.: Scaling in nature: from DNA through heartbeats to weather. Physica *A* 273 (1999) 46-69
7. Jenison, R.L., Fissel K.: A spherical basis function neural network for modeling auditory space. Neural Computation 8(1) (1996) 115-128
8. Slocum, A.C., Downey, D.C., Beer, R.D.: Further experiments in the evolution of minimally cognitive behavior: From perceiving affordances to selective attention. In: Meyer, J., Berthoz, A., Floreano, D., Roitblat, H., Wilson, S. (eds.): From Animals to Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior (2000) 430-439
9. Spronck, P.H.M., Sprinkhuizen-Kuyper, I.G., Postma, E.O.: Evolutionary Learning of a Neural Robot Controller. Proceedings of the International Conference on Computational Intelligence for Modelling, Control, and Automation (2001) 510-518
10. Stanley, H.E., Amaral, L.A.N., Gopikrishnan, P., Ivanov, P.Ch., Keitt, T.H., Plerou V.: Scale invariance and universality: organizing principles in complex systems. Physica A 281 (2001) 60-68
11. Viswanathan, G.M., Buldyrev, S.V., Havlin, S., da Luz, M.G.E., Raposo, E.P., Stanley H.E.: Optimizing the success of random searches. Nature 401 (1999) 911-914
12. Viswanathan, G.M., Afanasyev, V., Buldyrev, S.V., Havlin, S., da Luz, M.G.E., Raposo, E.P., Stanley, H.E.: Lévy flights search pattern of biological organisms. Physica A 295 (2001) 85-88

Panoramic View Based Monte Carlo Self-localization for Mobile Robots Operating in Real-world Environments

H.-M. Gross, H.-J. Boehme, Ch. Schroeter, and A. Koenig

Ilmenau Technical University, Department of Neuroinformatics, Germany
Horst-Michael.Gross@tu-ilmenau.de

Abstract. We present a novel panoramic view based robot localization approach which utilizes the Monte Carlo Localization (MCL) [1], a Bayesian filtering technique based on a discrete density representation by means of particles. We show how omidirectional imaging can be combined with the MCL-algorithm to globally localize and track a mobile robot given a taught graph-based representation of the operation area. To demonstrate the reliability of our approach, we present promising experimental results in the context of a challenging robotics application, the self-localization of a mobile service robot acting as shopping assistant in a very regularly structured, maze-like and crowded environment, a home store.

1 Introduction and motivation

Self-localization is the task of estimating the pose (position and orientation) of a mobile robot given a map of the environment and a history of sensor readings and executed actions. This includes both the ability of globally localizing the robot from scratch, as well as tracking the robot's position once its location is known. The localization problem is one of the fundamental problems in mobile robot navigation and many solutions have been presented in the past including approaches employing Kalman filtering, grid-based Markov localization, or Monte Carlo Methods [3]. The current state-of-the-art localization methods often use laser range finders or sonar, but these sensor modalities tend to be easily confused in environments with very regular topology, e.g. a supermarket or a home store with a great number of hallways of equal width, length and geometrical structure. Because of this maze-like topology, self-localization methods based on laser or sonar can produce numerous ambiguities complicating or preventing a quick self-localization or re-localization in case of a complete loss of positioning. In contrast, vision-based systems do not show these limitations, but supply a much greater wealth of information about the 3D-structure of the hallways and racks. For example, the filling of the goods racks gives the hallways a characteristic appearance, especially with respect to color or texture. Because of this, we expected to defuse the localization problem drastically by development of an approach for view-based localization that combines omnidirectional imaging with the probabilistic Monte Carlo Localization (MCL) [1].

2 Omnivision-based MCL

The Monte Carlo Localization (MCL) method underlying our omnivision-based localization approach is a version of Markov localization [6], a family of probabilistic approaches for approximating a multi-modal density distribution coding the robot's belief

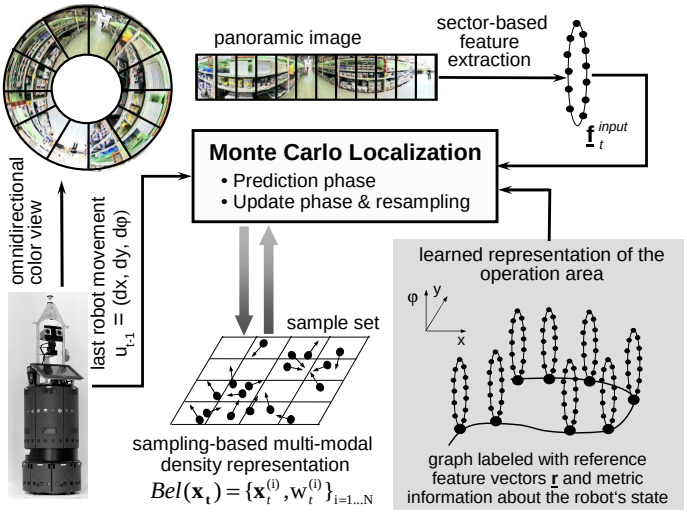


Figure 1. General idea of our omniview-based Monte Carlo Localization. The approach is based on a graph-based representation of the operation area. The nodes of the graph are labeled with both view-based visual features and metric information about the pose of the robot (position and heading direction in a world-centered reference frame) at the moment of the node insertion.

$Bel(\underline{\mathbf{x}}_t)$ for being in state $\underline{\mathbf{x}}_t = (x, y, \varphi)_t$ in its state space. x and y are the robot's position coordinates in a world-centered Cartesian reference frame, and φ is the robot's heading direction. The key idea of MCL is to represent the belief $Bel(\underline{\mathbf{x}}_t)$ by a set S_t of N weighted samples distributed according to $Bel(\underline{\mathbf{x}}_t)$: $S_t = \{\underline{\mathbf{x}}_t^{(i)}, w_t^{(i)}\}_{i=1..N}$. Here each $\underline{\mathbf{x}}_t^{(i)}$ is a sample, and the $w_t^{(i)}$ are non-negative numerical weighting factors called importance factors. Because the sample set constitutes a discrete approximation of the continuous density distribution, the MCL approach is computationally efficient, it places computation just "where needed".

The general idea of our view-based Monte Carlo Localization is illustrated in Fig. 1. In our approach, we use a graph-based representation of the operation area by a set of visual reference vectors $\underline{\mathbf{r}}(x, y, \varphi)$ extracted from the respective panoramic views at positions x, y in heading direction φ (Fig. 1, bottom right). The graph is constructed on-the-fly when manually joy-sticking the robot through the hallways of the store. During this training, omnidirectional images are captured from the environment and associated with the corresponding locations. For this purpose, in addition to the feature vectors extracted from the omnidirectional images, the nodes of the graph are labeled with metric information about the pose $\underline{\mathbf{x}} = (x, y, \varphi)$ of the robot at the moment of the node insertion. A new node (reference point) with importance for the representation is inserted, either if the Euclidian position distance to other reference points in a local Ω-vicinity or if the Euclidian feature distance between the current feature vector $\underline{\mathbf{f}}_t^{input}$ and the feature vectors $\underline{\mathbf{r}}_\Omega(x, y, \varphi)$ of these reference points are larger than given values. However, the labeling of the graph nodes with odometric data about the pose of the robot necessitates an efficient correction of odometry because of the increasing error over time, especially concerning the orientation angle. To attenuate this effect, we utilize a

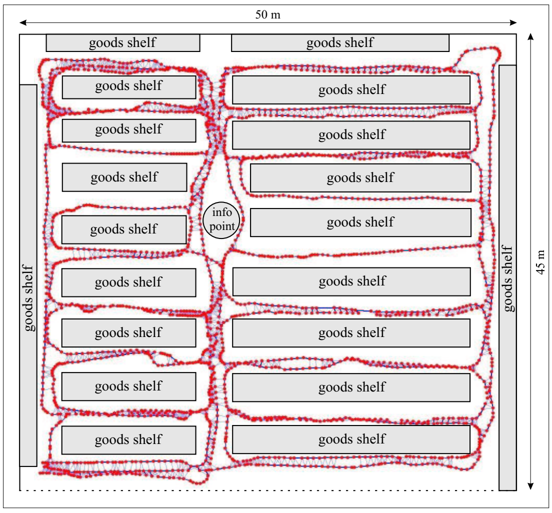


Figure 2. Topological map of the operation area in the home store. The size of the area is 50×45 meters, the graph consists of 2007 reference points (marked as dots) labeled with visual feature vectors and odometric data about the pose (position and orientation) of the robot at the moment of node insertion. The total distance travelled to learn this map was about 1000 meters.

specific feature of our market floor that shows a very regular structure caused by tiles that are uniquely oriented across the whole market area. For details of our vision-based odometry correction see [2]. We utilized this odometry correction method for learning a large-scale graph representation of the operation area as shown in Fig. 2 and achieved a very small absolute position error of about $60cm$ after a total distance of 1000 meters.

Feature extraction: Both during map-building and self-localization, the omnidirectional image is transformed into a panoramic image (see Fig. 1, top). Each panoramic image is first partitioned into a fixed number of non-overlapping sectors (typ. 10) each covering a part of the panoramic field of view. The following criteria determined the selection of appropriate features to describe the present scene: 1) To allow for an on-line localization, the calculation of the features should be as easy and efficient as possible. 2) The features should include the orientation of the robot as prerequisite to estimate the heading direction of the robot. 3) The feature description should allow for an easy generation of expected observations for unknown positions and orientations of the robot. 4) The features should be largely insensitive against partial occlusion of the environment, such as caused by people in the vicinity of the robot. Considering these criteria and the requirements of other omnivision-based localization approaches published recently, e.g. [4, 5], we decided to implement the simplest feature extraction method possible. Thereto, for each sector of the panoramic image, the mean RGB-color value is determined. This way, for each node in the graph a reference feature vector $\underline{\mathbf{r}}(x, y, \varphi)$ consisting of a small number of mean RGB-values has to be learned.

The localization algorithm: In analogy to the MCL algorithm presented in [1], our omniview-based MCL proceeds in two phases: In the *Prediction phase (robot motion)*, the sample set computed in the previous iteration (or during random initialization) is

moved according to the last movement of the robot u_{t-1} (Fig. 1, left). The *motion model* $p(x_t|x_{t-1}, u_{t-1})$ describes how the position of the samples changes using information u_{t-1} from odometry. This way, MCL generates N new samples that approximate the expected density distribution of the robot's pose after the movement u_{t-1}. To determine the expected observations $\underline{\mathbf{f}}_t^{(i)}$ of the moved samples, our approach requires interpolations both in state and feature space because of the coarse graph representation and the chosen feature coding. For each sample $s^{(i)}$, we first interpolate linearly between the reference feature vectors $\underline{\mathbf{r}}(x, y, \varphi)$ of the two reference nodes closest to the respective sample position $\underline{\mathbf{x}}_t^{(i)}$. After this, the resulting feature vector is rotated according to the expected new orientation $\varphi_t^{(i)}$ of the sample $s^{(i)}$. Since the feature vector only has a discrete number of components, we utilize a linear interpolation between the features of adjacent segments. This way, we obtain a set of N new feature vectors $\underline{\mathbf{f}}_t^{(i)}(x, y, \varphi)$ describing the expected observations of the moved samples in the new states $\underline{\mathbf{x}}_t^{(i)}$.

In the *Update phase (new observation)*, the actual panoramic view at the new robot position has to be taken into account in order to correct the sample set S_t. For this, the importance factor $w_t^{(i)}$ of each sample $s^{(i)}$ is computed. It describes the probability that the robot is located in the state $\underline{\mathbf{x}}_t^{(i)}$ of the sample. We determine the similarity $E_t^{(i)}$ between the current input feature vector $\underline{\mathbf{f}}_t^{input}$ extracted from the panoramic view at the new robot position and the expected feature vector $\underline{\mathbf{f}}_t^{(i)}$ of each sample $s^{(i)}$ simply by computing the angle between both normalized vectors applying a simple Gaussian-like *observation model*. Now $w_t^{(i)} = 1 - \alpha E_t^{(i)}$ can be determined, where α is a normalization constant that enforces $\sum_{j=1}^{N} w_t^{(j)} = 1$. The final sample set S_t for the next iteration is obtained by *re-sampling* from this weighted set. The re-sampling selects those samples with higher probability that have a high importance factor $w_t^{(i)}$. Samples with low importance factors are removed and randomly placed in the state-neighborhood of samples with high factors. After that, both phases are repeated recursively.

3 Experimental results

All experiments were carried out in the 'toom' home store Erfurt with our experimental platform PERSES, a standard B21 robot additionally equipped with an omnidirectional imaging system for vision-based navigation and human-robot interaction. The experiments were performed as off-line cross-validation tests on different sequences of images acquired in the home store. All images were labeled with the corresponding correct pose of the robot. One of the sequences is used as training data to build the graph while the other ones are used as test data (5000 pose-labeled images) to determine the localization error. Every localization experiment has a typical length of 190 movements, this corresponds to a path length of about 130 meters. Per experiment, the mean absolute localization error is determined. Every experiment was repeated 20 times, and the localization errors were averaged. It is to note that, in all cases, we studied the worst-case scenario: our robot had no prior information about its initial pose - this is a typical global localization problem. All tests can be judged as being very successful, as our localization system was able to find and continually track the position of the robot. Fig. 3 illustrates the typical course of a view-based self-localization and position

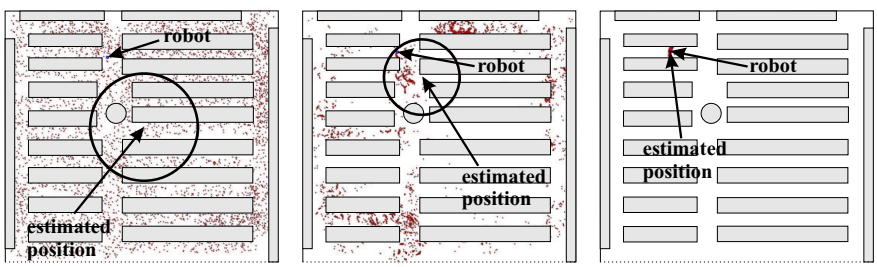


Figure 3. Self-localization and tracking experiment executed in a large section ($50 \times 45m^2$) of the home store. The sequence depicts the temporal condensation dynamics of about 4.000 samples (initial distribution, after 3 steps, and 9 steps). In the beginning, the robot is globally uncertain, the particles are spread uniformly throughout the free space. The variance of the 10% of the samples with the highest importance factors is marked as circle. Already after 9 movements (about 5,50 m), MCL has disambiguated the robot's position - the majority of samples is now centered tightly around the correct position, the variance is drastically reduced.

tracking experiment executed in a large section of the store ($50 \times 45m^2$). Despite the geometrical uniformity of the selected hallways and the coarse graph-structure (2007 nodes), our omniview-based MCL yields very precise localization results already after a few robot movements. For example, after 9 movements and observations, which corresponds to a travelled distance of about 5,50 meters, the difference between estimated and correct position of the robot was lower than 40 cm. The mean localization error of our test set is even smaller than 25 cm. The time required for computation of the MCL algorithm directly depends on the total number of samples. With the current on-board equipment (1500 MHz AMD Athlon), our algorithm requires about 50 ms for 4.000 samples. The time for image transformation and feature extraction takes about 25 ms per image. Therefore, our localization system enables real-time localization leaving a good amount of processing time for other navigation modules.

Dealing with occlusions: It is clear that we have to cope with occlusions in the scene, such as, for example, people walking by or objects being moved around in the environment. However, due to its wide visual field, occlusion of the entire panoramic view becomes very unlikely. For example, in Fig. 4 the two people standing as close as possible to the robot occlude no more than 10% of the visual field. To test the robustness of the localization algorithm, the test images were occluded by artificial gray-colored segments. The impact of occlusion effects was gradually controlled by the percentage of image content covered by the artificial image. Fig. 4 (bottom) depicts the results w.r.t. localization accuracy and various degrees of occlusion. For 0% occlusion, the mean position error is 25 cm and covers a range between 15 and 30 cm. The mean position error remains relatively low until 15% occlusion. Thereafter, the error vigorously increases since the image is affected by severe occlusions. However, due to the geometry of robot and vision system, it is not possible to place more than three or four people directly around the robot. Therefore, the maximum occlusion by people cannot be larger than 15-20%. Moreover, the internal particle dynamics of the MCL-algorithm realizes a kind of temporal self-stabilization of the estimation result, therefore, the influence of heavy but short occlusions can be largely neglected.

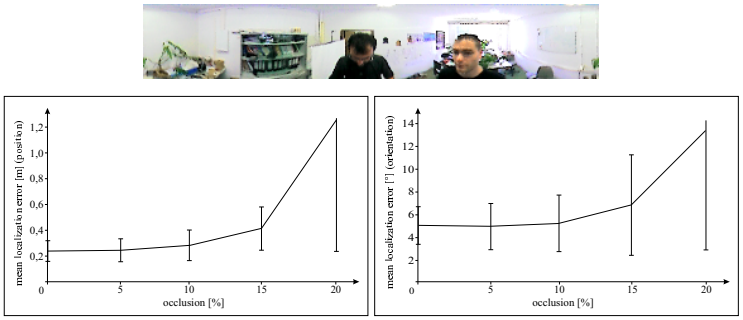


Figure 4. *(Top)* Occlusion example: two people are standing as close as possible to the robot and occlude about 10% of the visual field. *(Bottom)* Result of experiments investigating the influence of local occlusions on the position error *(left)* and and the orientation estimation *(right)*.

4 Conclusions and future work

In this paper, we have shown that particle filters in combination with a graph-based representation of the operation area by local panoramic views can be used to perform an omniview-based self-localization of a mobile robot in a challenging real-world application. Our localization system uses color omni-vision, works in real-time, and can easily be trained in new operation areas by joy-sticking. The results of the executed experiments confirm the accuracy and robustness of our omniview-based self-localization method.

Currently, theoretical and experimental studies are carried out to further improve our omniview-based MCL-system. For example, we are investigating the impact of the motion and observation models on the pose estimation and are studying the influence of a new mechanism adaptively controlling the sample rate on-the-fly on the localization accuracy. Other running experiments are dealing with the impact of appearance variations at the reference points in the learned graph, e.g. as result of a changed filling of the goods racks or modifications in the market topology. Moreover, our algorithm has to demonstrate its capabilities scaling up to the whole market area with a size of $100 \times 60m^2$ over a longer period of operation.

References

1. D. Fox, W. Burgard, F. Dellaert, S. Thrun. Monte Carlo Localization: Efficient Position Estimation for Mobile Robots. In: *Proc. AAAI-99*, 1999
2. H.-M. Gross, H.-J. Boehme. PERSES - a Vision-based Interactive Mobile Shopping Assistant. in: *Proc. IEEE Intern. Conf. on Systems, Man and Cybernetics, 2000*, pp. 80-85
3. J.-S. Gutmann, D. Fox. An Experimental Comparison of Localization Methods Continued. to appear: *Proc. IROS 2002*
4. B. Kroese, N. Vlassis, R. Bunschoten and Y. Motomura. A probabilistic model for appearance-based robot localization. *Image & Vision Computing*, 19 (6) 381-391, 2001
5. L. Paletta, S. Frintrop, and J. Hertzberg. Robust Localization Using Context in Omnidirectional Imaging. *Proc. ICRA 2001*, pp. 2072-2077, 2001.
6. S. Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artif. Intell.*, 99 (1998) 21-71

Visuomotor Adaptation: Dependency on Motion Trajectory

Christian Kaernbach[1], Lutz Munka[1], and Douglas Cunningham[2]

[1] Institut für Allgemeine Psychologie, Universität Leipzig
[2] Max-Planck Institut für Biologische Kybernetik, Tübingen
Christian@Kaernbach.de

Abstract. The present contribution studies the rapid adaptation process of the visuomotor system to optical transformations (here: shifting the image horizontally via prism goggles). It is generally believed that this adaptation consists primarily of recalibrating the transformation between visual and proprioceptive perception. According to such a purely perceptual account of adaptation, the exact path used to reach the object should not be important. If, however, it is the transformation from perception to action that is being altered, then the adaptation should depend on the motion trajectory. In experiments with a variety of different motion trajectories we show that visuomotor adaptation is not merely a perceptual recalibration. The structure of the motion (starting position, trajectory, end position) plays a central role, and even the weight load seems to be important. These results have strong implications for all models of visuomotor adaptation.

1 Introduction

In order to pick up an object, its visual location must be converted into the appropriate motor commands. Introducing an optical transformation (e.g., shifting the image horizontally via prism goggles) initially impairs this ability. The visuomotor system rapidly adapts to the discrepancy, however, returning performance to near normal.

von Helmholtz (1867), who was among the first to describe prism adaptation, reported that if one hand was active during adaptation, the other hand would also show an adaptation effect. It has by now often been demonstrated that intermanual transfer of adaptation is either very small or non-existent[1]. It is really quite striking that both hands have to adapt independently from each other. Consequently, prism adaptation can not be fully explained by "recalibrating" only visual perception so as to represent the seen location of an object correctly in spite of the prism goggles. However, this does not rule out a purely perceptual account of adaptation: the recalibration could

[1] Some studies (e.g. Choe and Welch, 1974) report intermanual transfer of adaptation. It is not clear, however, in how far this might be due to cognitive strategies. If participants are either ignorant of the effect of the goggles or repeatedly instructed to base their actions on their actual perception and not on cognitive strategies, intermanual transfer of adaptation is generally absent.

affect the proprioceptive perception of spatial location, i.e. the felt position of the arm. The proprioception of the active limb would have adapted while the proprioception of the passive limb would show no adaptation effects.

This notion of "perceptual learning" (e.g. Bedford, 1999) is seductive. As long as it is only the perceptual input that is recalibrated it is conceivable that spatial knowledge is represented centrally, in a kind of master data base, with all sensory systems providing calibrated spatial information. This data base would then in turn serve to provide the motor scripts with coordinate information of the objects that are to be dealt with. The performance difference for the active and the passive limb would be due to the different calibration status of the proprioceptive input to the central spatial representation from these limbs.

A central representation of spatial knowledge agrees well with the introspectively felt unity of phenomenal experience. However, it has been shown that phenomenal experience is not prerequisite for correct visuomotor behavior. Stratton (1897) has shown that wearing inverting goggles (turning the image 180°) perfect visuomotor coordination could be obtained within a few days. Phenomenally, however, the world was still upside down. It is still a matter of debate whether after a week or two phenomenal experience would also adapt; the important point here is that there is evidence for a dissociation between visuomotor and phenomenal adaptation. Comparable results were reported by Kohler (1951). On a similar line of thought evidence from blindsight cases (Pöppel et al., 1973) put into question the relevance of phenomenal experience for visuomotor functioning.

If visuomotor adaptation depends not only on the (active versus passive) limb but also on the exact motor trajectory, then a central representation of spatial knowledge would be less tenable. Instead, spatial knowledge would then be more easily and parsimoniously explained as distributed knowledge, closely related to a variety of possible motor scripts. Some initial evidence for such a dependency comes from Martin et al. (1996) who demonstrated that there was no transfer of adaptation from underhand to overhand throwing. Here, we examine this effect with the well-studied pointing task, as well as with types of movements that are more closely related than underhand and overhand throwing.

2 Experiment 1: Reaching Below/Above a Bar

Instead of measuring the adaptation effect directly, it has become common practice to measure the Negative Aftereffect (NAE), comparing motor performance before and after adaptation to prism goggles. It represents an excellent measure of adaptation as it compares two absolutely identical situations (unaltered vision) so that all observable changes in motor performance can only be due to the adaptation to the prism goggles that occurred in the meantime.

In Experiment 1 we measured the NAE for two different types of trajectories: Participants (N=72) had to touch a cross presented at eye level on a touch screen 30 cm in front of them. Two different trajectories were possible: reaching to the cross from below or (swinging the arm backwards) from above the horizontally extending bar that served as chin rest (Fig. 1). Location performance without feedback was deter-

mined for both trajectories of both hands before and after adaptation of a single trajectory of one hand to prism goggles (17° horizontal displacement). Testing was done on centrally located targets, while adaptation took place at horizontally displaced targets.

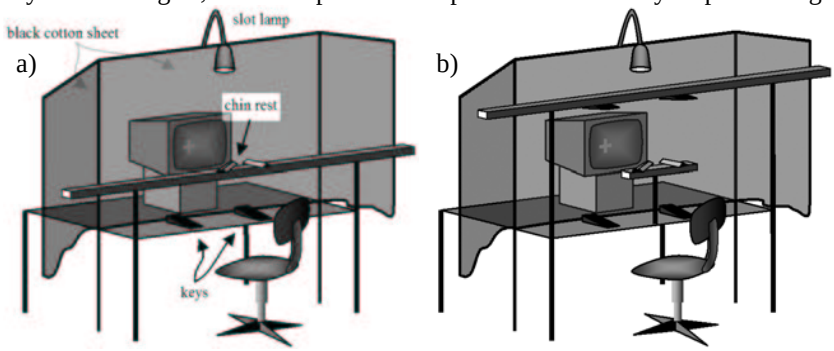


Fig. 1. a) Setup for Experiment 1. A thin black cotton sheet was hung in front of the touch screen, and the room light was shut off, with a dim slot lamp illuminating the hand while touching the screen. This procedure prevented additional visual cues whilst not hampering touching the screen or seeing the bright cross. b) The horizontally extending bar that served as chin rest was reduced in size for Experiments 2 to 6 so as to support the chin without hampering the more sweeping motions of those experiments. The top bar was used in Experiment 2.

The NAE was compatible with zero for both trajectories of the unadapted hand. This confirms the well-known finding that there is no intermanual transfer. More importantly, the NAE was significantly different for the adapted trajectory (46mm, 8.7°) as compared to the other trajectory of the same hand (26mm, $p<0.01$). That is, despite the fact that the starting positions were identical and end positions very similar, there was only partial *intramanual* transfer. The fact that there was partial transfer, rather than the complete absence of transfer found with overhand versus underhand throwing, reflects the greater similarity of the motions used here.

3 Experiment 2: Pointing from Different Starting Positions

While in Experiment 1 the starting position was identical for both types of trajectories, the end positions were slightly different. In order to exclude the possibility that this caused the weak intramanual transfer, Experiment 2 was run using different starting positions and identical end positions. The setup differed from that of Experiment 1 in that the chin rest did not extend horizontally, and there was a horizontal bar mounted 90 cm above the table, with two additional keys mounted beneath that bar. Participants (N=21) performed a total of 45 sessions, starting the pointing movement either at a low (desktop key) or a high position (key mounted beneath top bar). Location performance without feedback was determined for both starting positions before and after adaptation to a single starting position while wearing prism goggles. – The NAE was again significantly different for the adapted starting position (80 mm) as compared to the other starting position (51 mm, $p<0.01$).

4 Experiment 3: Interposing Inward/Outward Circles

In Experiments 1 and 2, either the starting positions or the end positions differed. In Experiment 3, participants (N=14, performing a total of 32 sessions) started the pointing movement at the same position (at the key on the desk top), and ended it with the same end position. Instead of moving their hand directly from the key to the cross, they had to interpose an inward or outward circular movement. They were instructed to circumscribe a region "the size of a head", like writing a kind of "O" in the air, after releasing the key, and before touching the screen. Location performance without feedback was determined for both trajectories before and after adaptation to a single trajectory while wearing prism goggles. – Even with identical starting and end positions, the NAE was significantly different for the adapted trajectory (59 mm) as compared to the other trajectory (49 mm, $p<0.01$). The difference is, however, smaller than in Experiments 1 and 2: The NAE for the two trajectories differed by 17%, whereas the difference was around 40% in the other two experiments.

5 Experiment 4: Pointing with/without a Weighted Wristband

In Experiments 1 to 3, trajectories differed. In Experiment 4, transfer of adaptation was studied for *exactly the same trajectory*, varying this time the load of the moving arm by applying a weighted wrist band (440 g) in some of the trials. Participants (N=11) performed a total of 36 sessions. Again, location performance without feedback was determined for both conditions before and after adaptation to a single condition while wearing prism goggles. – Varying only the load of the moving arm, the NAE was again significantly different for the adapted condition (55 mm) as compared to the other condition (44 mm, $p<0.05$). The NAEs differed by about 22%.

6 Experiment 5: Generalization to Vertically Distributed Targets

In Experiments 1 to 4, we made use of the fact that adaptation generalizes horizontally: Participants adapted to targets that were to the side of the centrally located targets used in the pre- and post-tests (see methods of Experiment 1). Generalization of adaptation horizontally to other targets has been demonstrated before (Bedford, 1993). As prism goggles displace the image horizontally, this is not too surprising. By the same token, it is not necessarily clear that adaptation will generalization vertically. In Experiment 5, participants (N=14, performing 20 sessions) adapted to a high target position, or to a low position, or alternately to both these positions. Their location performance before and after adaptation was tested at a high, a medium and a low target position. In order to obtain a good separation (30 cm, corresponding to 53° visual angle) between high and low target positions, the monitor was rotated 90°.

Figure 2 shows the results. When adapting to the high target position, this condition showed the largest NAE, with a gradual decrease of the NAE as the tested position departs from the adapted one. The differences between testing the high target

position and the other two target positions is significant ($p<0.05$). The same trend is present when adapting to the lower target position, although this trend did not reach significance. When adapting alternately to both high and low target positions, no significant differences are to be found.

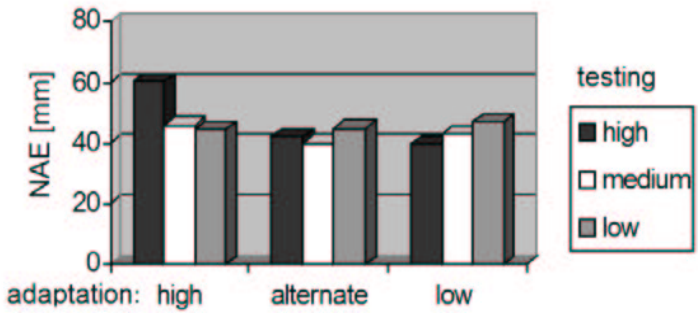


Fig. 2. Results of Experiment 5. NAE as a function of adapted and tested target position.

As can be seen in Fig. 2, generalization for vertically distributed targets is not perfect. The effect is, however, too small to be evaluated well within the distances that can be realized on a rotated touch screen. Future experiments will include target positions outside the touch screen area.

7 Experiment 6: Effect of Terminal/Full Feedback

In a final experiment we assessed the effect of feedback and the speed of adaptation. In the previous experiments, adaptation took place under “full feedback”, i.e. the participants could watch their hand as it moved towards the target. Under full feedback, participants usually produce only small location errors, correcting errors of the ballistic part of the motion while approaching the screen. These data do not allow analysis of the dynamics of the adaptation process. In Experiment 6, participants (N=19, performing 28 sessions) adapted either alternately under full feedback and under no feedback (with the no-feedback trials yielding information on the state of the adaptation), or under “terminal feedback”: In this condition, the lamp went off when the participant released the key, and went on again when the screen was touched. We reasoned that under terminal feedback the participant would realize the true mistake of the ballistic movement which would be obscured under full feedback due to the possibility to correct the movement “on the fly”. We expected that terminal feedback would induce a stronger adaptation effect. – Figure 3 reveals that indeed terminal feedback induces a stronger NAE than full feedback ($p<0.01$). The initial adaptation speed seems not to be affected, but the final adaptation level is greater after terminal feedback.

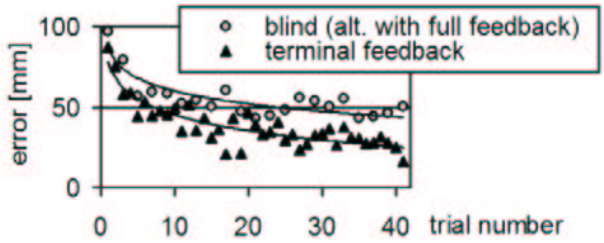


Fig. 3. Results of Experiment 6. Location error during adaptation session as a function of trial number and feedback condition.

Conclusion

Our data demonstrate that, in a variety of cases, spatial adaptation is motor specific: *Knowing where is knowing how to* (see also the reinterpretation of the *what* and *where* systems by Milner and Goodale, 1995). While spatial knowledge seems to be distributed, we nonetheless phenomenally experience it as a unitary entity. Even if "left arm knowledge" differs from "right arm knowledge" (due, e.g., to adaptation of one arm), we do not perceive any ambiguity when seeing an object. The cause of this dissociation might be elucidated by considering the purpose served by the experienced unity of spatial knowledge. Phenomenal experience is a late product of evolution, enabling the individual to plan coherent sequences of actions (and anticipate their consequences), as has e.g. been demonstrated with rats (Tolman, 1948). For such a purpose it would probably be cumbersome to be aware of the fragmentation of spatial knowledge, including possible inconsistencies. Simple aim-directed reactions to visual input (as in pointing or grasping) have developed earlier and are apparently implemented independently at a level closely related to motor performance.

References

Bedford, F. (1993). Perceptual Learning. In *The psychology of learning and motivation*, Vol. 30. D. Medin (Ed.). Academic Press, San Diego, CA, pp. 1-60

Choe, S. C., and Welch, R. B. (1974). Variables affecting the intermanual transfer and decay of prism adaptation. *Journal of Experimental Psychology*, 102, 1076-1084.

von Helmholtz, H. (1867). *Handbuch der physiologischen Optik.* Leipzig: Voss.

Kohler, I. (1951). Über Aufbau und Wandlungen der Wahrnehmungswelt. Österreichische Akademie der Wissenschaften. Sitzungsberichte, philosophisch-historische Klasse, 227, 1-118.

Martin, T.A., Keating, J.G., Goodkin, H.P., Bastian, A.J., and Thach, W.T. (1996). Throwing while looking through prisms. II. Specificity and storage of multiple gaze-throw calibrations. *Brain*, 119, 1199-1211.

Milner, D. and Goodale, M., (1995). *The Visual Brain in Action*, Oxford University Press.

Pöppel, E. , Held, R. and Frost, D. (1973). Residual function after brain wounds involving the central visual pathways in man. *Nature*, 243, 295-96.

Stratton, G. (1897).Vision without inversion of the retinal image. *Psychological Review*, 4, 361-360.

Tolman, E.C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55, 189-208.

Detection of Communication Partners from a Mobile Robot

S. Lang, M. Kleinehagenbrock, J. Fritsch, G. A. Fink, and G. Sagerer

Bielefeld University, Faculty of Technology, 33594 Bielefeld, Germany
slang@techfak.uni-bielefeld.de

Abstract. An important prerequisite for the natural interaction of humans with a mobile robot is the robot's capability to detect potential communication partners. In this paper we present an approach which uses a combination of person recognition and tracking with sound source localization realized in a multi-modal anchoring framework. As in open environments several potential communication partners can be present simultaneously we developed a rule-based method for selecting one specific person as the current communication partner.

1 Introduction

A prerequisite for the widespread use of mobile service robots in home and office environments is the development of systems with natural human-robot-interaction. While much research focuses on the communication process itself, it is also necessary to explore how robots can automatically recognize when and how long a user's attention is directed towards the robot for communication.

For this purpose some fundamental abilities of the robot are required. It must be able to detect persons in its vicinity and to track their movements over time. Additionally, as speech is the most important means of communication for humans, the detection and localization of sound sources is of great importance.

This paper is organized as follows: At first we discuss approaches that are related to the detection of communication partners in section 2. Then, in section 3 multi-modal anchoring is described. This is the basis of our approach for the detection of communication partners explained in section 4. The paper concludes with a short summary.

2 Related Work

As long as artificial systems interact with humans in static setups the detection of communication partners (CPs) can be achieved rather easily. For the interaction with an information kiosk the potential user has to enter a well defined space in front of the device (cf. e.g. [1]). In intelligent rooms usually the configuration of the sensors allows to monitor all persons involved in a meeting simultaneously (cf. e.g. [2]).

* This work has been supported by the German Research Foundation within the Collaborative Research Center 'Situated Artificial Communicators' and the Graduate Programs 'Task Oriented Communication' and 'Strategies and Optimization of Behavior'.

In contrast to these scenarios a mobile robot does not act in a closed or even controlled environment. A prototypical application of such a system is its use as a tour guide in scientific laboratories or museums (cf. e.g. [3]). All humans approaching or passing the robot have to be considered to be potential CPs. In order to circumvent the problem of detecting humans in an unstructured and potentially changing environment in [3] a button on the robot itself has to be pushed to start the interaction.

The humanoid robots *SIG* [4] and *ROBITA* [5] currently demonstrate their capabilities in research labs. Both use a combination of visual face recognition and sound source localization for the detection of potential CPs. *SIG*'s focus of attention is directed towards the person currently speaking that is either approaching the robot or standing close to it. In addition to the detection of talking people *ROBITA* is also able to determine the addressee of spoken utterances. Thus it can distinguish speech directed towards itself from utterances spoken to another person. Both robots, *SIG* and *ROBITA*, can give feedback which person is currently considered to be the CP. *SIG* always turns its complete body towards the CP. *ROBITA* can use several combinations of body orientation, head orientation, and eye gaze to express different states of communication.

3 Anchoring

Person tracking with a mobile robot is a highly dynamic task. Due to motions of the tracked persons and of the robot itself the sensory perception of the persons is constantly changing. In order to control the robots behavior, connections between processes that work on the level of abstract representations of objects in the world (symbolic level) and processes that are responsible for the physical observation of these objects (sensory level) need to be established. These connections, called *anchors*, must be dynamic, since the same symbol must be connected to new percepts every time a new observation of the corresponding object is acquired.

We follow the definition of anchoring proposed in [6]: At every time step t, the anchor contains three elements: a symbol, which is used to denote an object, a percept of the same object, generated in the perceptual system, and a signature, meant to provide the estimate of the values of the observable properties of the object. If the anchor is grounded at time t, it contains the percept perceived at t as well as the updated signature. If the object is not observable at t and therefore the anchor is ungrounded, then no percept is stored in the anchor but the signature still contains the best available estimate.

3.1 Multi-Modal Anchoring

Anchoring as defined in [6] only considers the special case of connecting one symbol to the percepts acquired from one sensor. However, complex objects cannot be captured completely by a single sensor system alone. If more than one sensor is used, the symbolic description of an object has to be linked to different types of percepts, originating from different perceptual systems.

For this purpose we propose an approach for *multi-modal anchoring* [7]. It allows distributed anchoring of individual percepts from multiple modalities and copes with

different spatio-temporal properties of the individual percepts. Every part of the complex object which is captured by one sensor is anchored by a single *component anchoring process*. The composition of all component anchors is realized by a *composite anchoring process* which establishes the connection between the symbolic description of the complex object and the percepts from the individual sensors. In addition to standard anchoring, the composite anchoring module requires a *composition model*, a *motion model*, and a *fusion model*:

- The composition model defines the spatial relationships of the components with respect to the composite object. It is used in the component anchoring processes to anchor only those percepts that satisfy the composition model.
- The motion model describes the type of motion of the complex object, and therefore allows to predict its position. Using the spatial relationships of the composition model, the position of percepts can be predicted, too. This information is used by the component anchoring processes in two ways: 1. If multiple percepts were generated from one perceptual system the component anchoring process selects the percept which is closest to the predicted position. 2. If the corresponding perceptual system receives its data from a movable sensor with a limited field of view (e.g. pan-tilt camera), it turns the sensor into the direction of the predicted position.
- The fusion model defines how the perceptual data from the component anchors has to be combined. It is important to note, that the processing time of the different perceptual systems may differ significantly. In this case the perceptual data is not received by the composition anchoring process in chronological order. For this purpose the composite anchor provides a chronologically sorted list of the fused perceptual data. New data from the component anchors is inserted in the list, and all subsequent entries are updated.

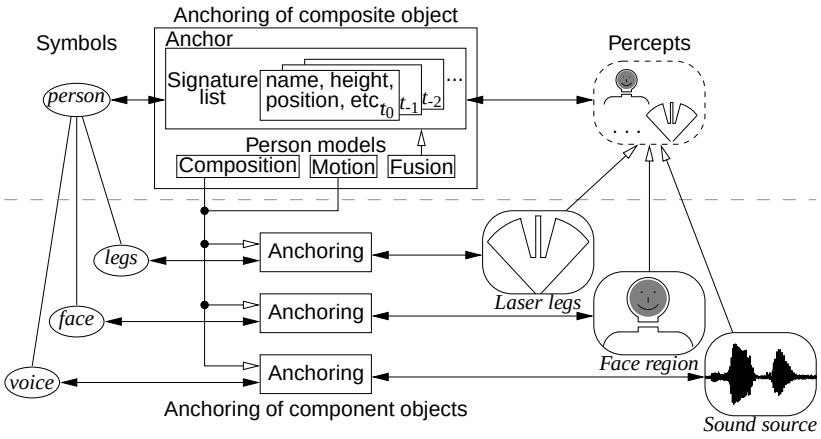


Fig. 1. Multi-modal anchoring of persons.

The detection of CPs from a mobile robot requires to track all persons in the vicinity of the robot. For this purpose we apply multi-modal anchoring. Our mobile robot is

equipped with a laser range finder, a pan-tilt camera, and stereo microphones. Every sensor forms the basis for one perceptual system:

- The laser range finder covers a 180° field of view at a height of approximately 30cm. In range readings human pairs of legs result in a characteristic pattern that can be easily detected [7]. From a *legs*-percept distance and angle of the person relative to the robot are extracted.
- The camera is mounted on top of the robot at a height of 140cm. We developed a face detection method which copes with changing lighting conditions [8]. From a *face*-percept the distance, angle, height and identity of the person are extracted.
- Stereo microphones are applied to locate speakers using a method based on cross-powerspectrum phase analysis [9]. From a *voice*-percept the angle relative to the robot can be extracted.

The anchoring of a person is illustrated in Fig. 1. It is based on anchoring the three components *legs*, *face*, and *voice*.

3.2 Anchoring Multiple Persons

For the detection of CPs from a mobile robot usually more than one person has to be tracked at the same time. Then, several anchoring processes have to be run in parallel. In this case, multi-modal anchoring as described in the previous section may lead to the following conflicts between the individual composite anchoring processes:

1. A percept is selected by more than one anchoring process.
2. The anchoring processes try to control the pan-tilt unit of the camera in a contradictory way.

To resolve these problems a *supervising module* is required, which controls the selection of percepts and the access to the pan-tilt camera.

To handle the first problem, the supervising module restricts the access to the pan-tilt unit of the camera to only one composite anchoring process at a time. How access is granted to the processes depends on the intended application. An example is given for the detection of CPs in the following section.

In order to avoid the second problem, the selection of percepts is implemented as follows. Instead of selecting a specific percept deterministically every component anchoring process assigns scores to all percepts rating the proximity to the predicted position. Subsequently, the supervising module computes the optimal non-contradictory assignment of percepts to component anchors. Percepts that are not assigned to any of the existing anchoring processes are used to establish new anchors. Additionally, an anchor that was not updated for a certain period of time will be removed by the supervising module.

4 Detection of Communication Partners

For the detection of CPs from a mobile system we apply multi-modal anchoring of persons, as described in the previous sections. Every person in the vicinity of the robot

is anchored by one anchoring process. From the anchoring processes the following attributes can be extracted: *standing*: The last positions of a person are known; it can therefore be decided, whether a person is walking or standing still. *speaking*: From the microphones it is known, whether a person is speaking or is silent. *facing*: The face anchoring process provides information, whether a person is facing the robot or looking in a different direction.

Note, that due to the limited field of view of the camera the attribute *facing* can not be computed for all persons simultaneously. The control of the access to the pan-tilt unit of the camera by the supervising module has an important relevance for this application.

For CP detection we propose the following set of heuristic rules, that are based on the three above mentioned attributes *standing*, *speaking*, and *facing*:

- A person that is not standing but walking is considered as a passer-by, and is therefore definitely no CP (rule 1).
- Whether a person standing still that is also speaking is classified as CP depends on the orientation of the head:
  - A person facing the robot is classified as CP (rule 2).
  - A person not facing the robot is assumed to be talking to someone else than the robot (e.g. another person) and therefore is definitely no CP (rule 3).
  - If no information from the camera is available, no classification is possible, so the person is a potential CP (rule 4).

The remaining three configurations of attributes (*standing*, not *speaking*, and any state of *facing*) leave the person's state unchanged (rule 5). This means that a person which was previously recognized as CP will be still considered as CP.

The rules for the detection of CPs are now used to define the behavior of the robot. On the one hand, the robot should direct its attention towards the person which was recognized as CP. This is done by turning the front of the robot into the direction of the CP, standing *face-to-face*. The anchoring process corresponding to that person gets access to the pan-tilt camera and keeps the person in the center of the field of view. On the other hand, the robot must be able to recognize a new CP, when the current CP is not speaking (rule 5). Only a person that is speaking can take over the role of the CP. If a person speaking is in the field of view of the camera one of the rules 2 or 3 can be applied and a decision is possible. If a person speaking is not in the field of the camera it is considered as a potential CP (rule 4). Then, the corresponding anchoring process gets access to the pan-tilt camera in order to focus the potential CP. Now a decision can be made according to rules 2 and 3. If the person is facing the robot it becomes the new CP, otherwise the anchoring process of the old CP again gets back access to the pan-tilt camera. Note, that while the camera is used to check the state of other persons the front of the robot is still directed towards the current CP, thus signaling that this person is the current CP. A sample behavior of the robot is depicted in Fig. 2.

5 Summary

We presented an approach for the detection of communication partners (CPs) from a mobile robot. The detection requires to simultaneously track persons in the vicinity of

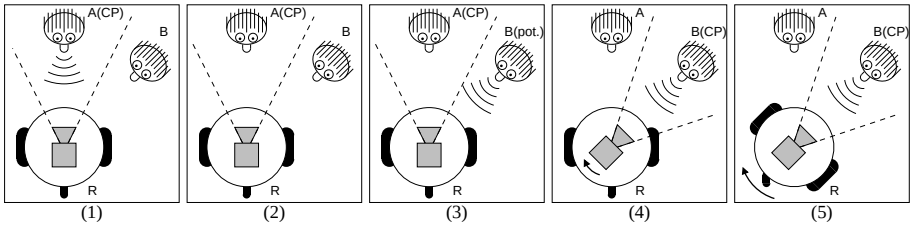


Fig. 2. Sample behavior with two persons standing near the robot R. In (1) person A is the CP, thus the robot directs its attention towards A. Then A stops speaking but remains the CP (2). In (3) person B begins to speak. Unless B's head is not in the camera's field of view, B is a potential CP. Therefore the robot turns the camera into the direction of B, still showing A its attention (4). Since person B is facing the robot, B becomes the new CP, and the robot turns towards B in (5).

the robot. This is achieved by multi-modal anchoring based on three types of sensors: pan-tilt camera, laser range finder, and stereo microphones. The anchoring processes provide three person attributes: standing, speaking, facing. We developed a set of heuristic rules which define if a person is considered as a CP. In addition, the competing access of the anchoring processes to the pan-tilt unit of the camera is described.

References

1. V. Pavlović, A. Garg, J. Rehg, and T. Huang. Multimodal speaker detection using error feedback dynamic bayesian networks. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR-00)*, pages 34–43, Los Alamitos, June 2000.
2. R. Stiefelhagen, J. Yang, and A. Waibel. Estimating focus of attention based on gaze and sound. In *Workshop on Perceptive User Interfaces (PUI'01)*, November 2001.
3. W. Burgard, A. B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. The interactive museum tour-guide robot. In *Proc. of the Fifteenth National Conf. on Artificial Intelligence (AAAI-98)*, 1998.
4. H. G. Okuno, K. Nakadai, and H. Kitano. Social interaction of humanoid robot based on audio-visual tracking. In *Proc. of 18th Intern. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-2001)*, June 2002.
5. Y. Matsusaka, S. Fujie, and T. Kobayashi. Modeling of conversational strategy for the robot participating in the group conversation. In *Proc. European Conf. on Speech Communication and Technology*, pages 2173–2176, Aalborg, Denmark, September 2001.
6. S. Coradeschi and A. Saffiotti. Perceptual anchoring of symbols for action. In *Proc. of the 17th IJCAI Conf.*, pages 407–412, Seattle, WA, 2001.
7. M. Kleinehagenbrock, S. Lang, J. Fritsch, F. Lömker, G. A. Fink, and G. Sagerer. Person tracking with a mobile robot based on multi-modal anchoring. In *IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, 2002. to appear.
8. J. Fritsch, S. Lang, M. Kleinehagenbrock, G. A. Fink, and G. Sagerer. Improving adaptive skin color segmentation by incorporating results from face detection. In *IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, 2002. to appear.
9. D. Giuliani, M. Omologo, and P. Svaizer. Talker localization and speech recognition using a microphone array and a cross-powerspectrum phase analysis. In *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1243–1246, Yokohama, Japan, 1994.

Real Time Object Recognition in a Dynamic Environment

An application for soccer playing robots

Tino Lourens and Emilia Barakova

GMD-Japan Research Laboratory
2-1 Hibikino, Wakamatsu-ku
Kitakyushu, 808-0135, Japan
http://www.gmd.gr.jp
{tino, emilia}@gmd.gr.jp

Abstract In dynamic environments, such as RoboCup [4], vision systems play a crucial role. In general, systems requiring real-time vision are either implemented in hardware, or as software systems that take advantage of the domain specific knowledge to attain the necessary efficiency. The goal of this paper is to describe a vision system that is able to reliably detect objects in real time and that is robust under different lighting conditions, this in contrast to most models used in robot soccer. The resulting objects serve as input for intelligent prediction of robot behavior [1].

1 Introduction

Fast sensing is advantageous for both biological and artificial systems. Humans can evaluate a visual scene in a fraction of a second. During this period a considerable amount of data is retrieved and processed. Humans very rapidly reduce visual data (the eyes receive most data per time unit) to relevant information, by using knowledge and adaptation. This paper describes a system where visual data is rapidly reduced by using knowledge about the environment. The system is able to recognize objects in real time, and is applied to soccer playing robots. The system is robust to (non-uniform and changing) lighting conditions and differences in color definition. This in contrast to virtually all vision solutions in RoboCup[1], which need tedious tuning for every game.

In most of the color vision systems the first step in data processing are extracting features, assigning pixels to classes, or a combination of both. In general, software solutions for object recognition are not even close to real time. For example, a well known fast method of feature (corners and edges) detection is SUSAN [5]. The newest generation of processing technology might be able to perform this method in real time. Nevertheless, the largest computational effort is needed to map the detected key-points into recognized objects. Hence,

[1] RoboCup is the robot soccer competition.

knowledge possibly in combination with selective attention is essential for real time vision applications, for general tasks in a real world environment.

Assigning pixels to classes has been proven to be successful in RoboCup. The most widely used approach in RoboCup is assigning pixels to color classes. Bruce et al. [2] constructed a thresholding method that is able to classify up to 32 different color classes in few operations. This method is attractive because of its efficient memory usage. We propose a simpler method that assigns a pixel to a class in a single operation by a lookup table.

Colored objects can be recognized as blobs under the assumption that objects are uniform. Bruce et al. [2] accomplished blob detection by color segmentation using run length encoding. The advantage of this method is that it is independent of any knowledge about the environment. The drawbacks of this method however include, poor recognition of partly occluded objects and the relatively high computational cost.

A more attractive approach is proposed by Jamzad et al. [3]. They make use of the perspective view, and state that 1200 single points are sufficient for object detection in RoboCup. In the future, robots are supposed to play against humans by the FIFA rules. The replacement of the boarding by white lines, last year, is a step in that direction. To meet the new requirements, there is a stronger need for detecting small objects, hence so-called scanlines are more suitable than single points.

The paper is organized as follows: Section 2 elaborates on color space reduction, spatial reduction, and cluster extraction to obtain real time object detection in a robot soccer playing environment. The paper concludes with a discussion and future research.

2 Object Recognition in a RoboCup Environment

Detection of objects for a soccer playing robot comprises of four stages: *Color space reduction*; in a RoboCup setting only few colors are relevant, therefore a natural image is reduced to these relevant colors in a single operation by using a preprocessed table. *Spatial reduction*. A perspective grid with a resolution of 10 cm reduces the spatial data to less than 20 percent, which is sufficient to recognize all objects. *Segmentation* is performed by clustering areas of identical color on the grid. *Object extraction*; the objects contain at most 3 different color clusters and are evaluated by size.

2.1 Color Space Reduction

The colors of all objects in RoboCup are defined. These colors are *orange* for the ball, *yellow* and *blue* for goals and corner poles, *green* for the field, *white* for the lines, *black* for the robots, and *cyan* and *magenta* to visually distinguish between teams. We omit cyan, because we are able to distinguish our robots from the others by team communication. We define these seven colors as *color classes* because they cover more than a single (r, g, b) color value.

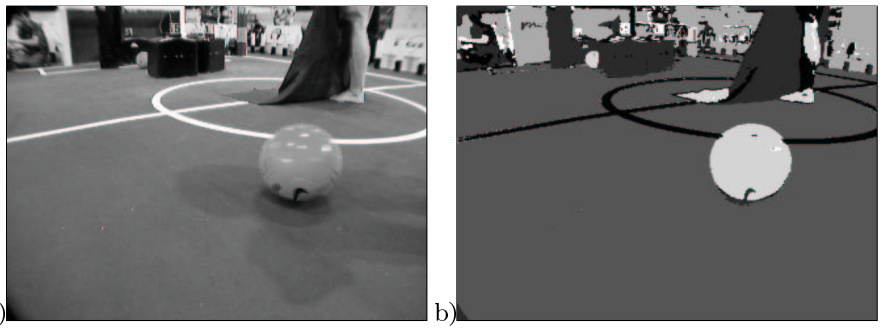


Figure 1. a) Input image. b) Results of color reduction by attaching every pixel to one of the seven color classes. Note that the image is displayed by a normalized index, to clearly differentiate between the seven classes.

Color space reduction is accomplished by assigning every pixel in an image to a single color class.[2] This reduction is achieved in a single operation by a preprocessed table of $N_1 \times N_2 \times N_3$ elements, where N_k denotes the number of colors in channel $k \in \{1, 2, 3\}$.

A color triple $c = (r, g, b)$ in the conversion table t is assigned to exactly one color class:

$$t(c) = \begin{cases} \text{white} & \text{if } (\max - \min) < U) \wedge (\text{avg} > T) \\ \text{black} & \text{if } (\max - \min) < U) \wedge (\text{avg} \leq T) \\ i & \text{if } (\max - \min) \geq U) \wedge (|c' - i| \leq |c' - j| \quad \forall j \in C) \end{cases} \tag{1}$$

where U is a uniformity measure, T is a threshold, $\text{avg} = |c|$ is the average, $\min = \min_3(c)$ is the minimum, $\max = \max_3(c)$ is the maximum element value in c. The normalized color $c' = (N_1(r - min)/max, N_2(g - min)/max, N_3(b - min)/max)$, $C = \{orange, blue, yellow, green, magenta\}$ as the set of "real" color classes, and $|x|$ denotes the Euclidian (or L_2) distance. The following settings are used throughout the paper: $U = 50$, $T = 100$, and $N_k = 255$ for every k; the color centers of a class are used in its exact definition (blue $= (0, 0, 255)$, orange $= (255, 165, 0)$, etc.). These settings turn out to be very robust, but can of course be set to the the most desired settings in a particular RoboCup environment. Figure 1 illustrates the results of this algorithm.

2.2 Processing Data in a Perspective View

The robots are equipped with a Sony DFW VL500 CCD camera and have attached a Sony wide angle lens (x0.6 VCL-0637H). The camera is connected to a

[2] The Kmeans algorithm can be used for finding the most appropriate colors in environments where time constraints are less critical and where color settings are not a-priori known.

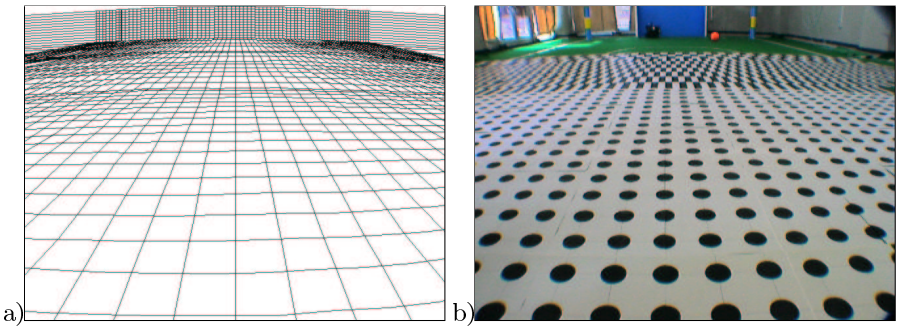


Figure 2. a) Constructed two-dimensional grid. b) Calibration pattern. Equidistant blobs, with centers at 10 cm distance frome each other, are used up to two meters. At a larger distance a 10x25 cm grating pattern is used.

notebook by the IEEE1394 firewire bus. The maximum capacity throughput of the camera is used, which results in 640x480 color (YUV422) images at a 30 Hz framerate.

In RoboCup soccer this setup is sufficient to detect a ball at a 10 meter distance. The image data can be strongly reduced if the sizes of all objects and their positions on the soccer field are taken into account. The smallest static object of interest in a soccer field is the white line (12 cm width) the next smallest object is the ball that has a diameter of about 24 cm.

A perspective grid (Figure 2a) is constructed by using a calibration pattern that contains equidistant blobs that are at 10 cm distance (Figure 2b). Such a grid highly reduces the data, and is still sufficient to recognize all objects in a RoboCup environment. The depthlines (7.08 percent) are completely scanned. Depending on the content of the depthlines, part of the 42 horizontal scanlines (8.75 percent) is scanned. In addition the grid gives the world coordinates from robot perspective (angle and depth to an object) which serve as input for the motion control as well as behavior prediction[1].

2.3 Object Extraction

The objects in a RoboCup setting are all uniform in color. However, in practice, differences in definition of color, reflecting surfaces, illumination of different light sources, as well as, light from outside can have a strong influence on the uniformity and appearance of a color. Actually, this is the major problem in RoboCup vision. The choice of assigning every pixel to an a-priori known small number of color classes (Section 2.1) is robust to these differences in appearance.

A grid (Figure 2a) is placed over the image with assigned color classes. All depthlines of the grid are used to scan for objects. On a single depthline all sequences of pixels that belong to the same color class and exceed a fixed minimum length are evaluated. The minimum and maximum y-coordinate of a sequence

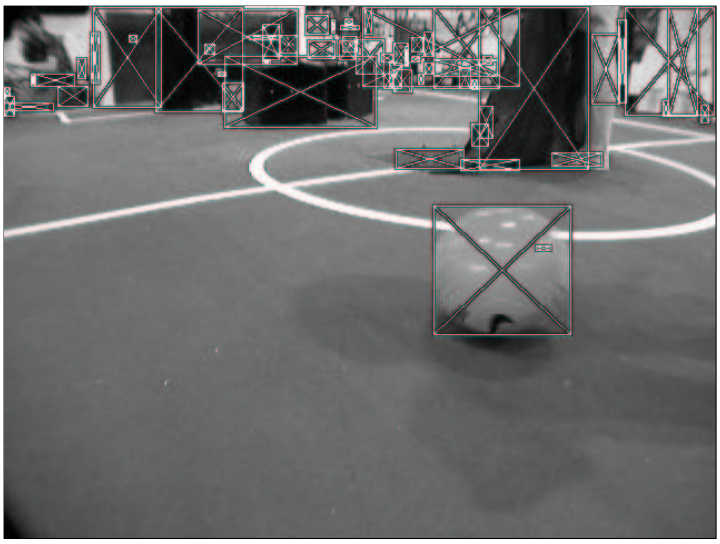


Figure 3. Extracted color clusters of Figure 1a are marked by a rectangle with a cross. Five different color clusters (orange, blue, yellow, black, and magenta) are evaluated. Unfortunately they can not be displayed properly in a grey scale image.

are taken and denote the vertical size of a cluster. Next, all intersections of the sequence with the horizontal scanlines are marked and followed in left and right direction, until another color class is encountered. The maximum and minimum x-coordinates of the followed horizontal scanlines determine the horizontal size of the cluster. If this cluster intersects with an existing cluster it is merged, otherwise a new cluster is added to the set of clusters in a single (time) frame. An illustration of all marked clusters of Figure 1a is given in Figure 3.

When all depthlines are followed in a single frame, the set of segments is complete and object recognition is performed. An object is described by a set of connected clusters and by its size. For example, a corner pole is a blue-yellow-blue object, with a diameter of about 20 cm and a height of about one meter; a ball consists of a single color cluster of about 24 times 24 cm.

2.4 Results

The proposed real time object recognition system is bound by the capacities of the video camera. The load on a 850 MHz P3 notebook is around 90 percent and that of a 2.2 GHz P4 desktop PC is around 35 percent. The initialization needed for the construction of color mapping table, grid, depth map, and map for size estimation are all fully preprocessed, which requires between 10 to 20 seconds, depending on the used machine.

Grabbing and mapping the data into a color segmented image, which can be considered as data retrieval, is most time consuming (240 ms wall clock

time for 30 frames on the P4 desktop). Object recognition itself is far less time consuming (102 $\pm$ 5 ms). These measurements are obtained from data taken from two different RoboCup environments. In all cases the number of extracted color clusters are between 20 and 80 in a single frame.

3 Discussion and Future Research

In this paper a real time object recognition model in a RoboCup environment that is robust under varying illumination conditions and differences in color definition is presented. The model includes color and spatial reduction, by assigning pixels to color classes and by using a grid, respectively.

The method provides simple incorporation of additional color classes and allows segments to have more than one color class. For example, between the yellow and orange class a few intermediate color classes can be defined to give a more accurate distinction between a yellow goal and an orange ball. The green color is currently ignored, but segments very well from all other color classes and can be used to determine the field boundaries.

The current setup contains a vision system with a wide angle lens. Omnidirectional vision which is commonly used in RoboCup suffices in resolution. The 360 degree field of view results in simpler hardware and better self localization, and will be used on a robot that is under development. In the model only the grid needs to be recalibrated.

In autonomous systems where vision is included there are three major categorized data streams: *color*, *form*, and *motion*. These three streams are essential for any basic vision system. In real world applications that are not predefined like RoboCup, will require visual attention, knowledge, and learning methodologies to quickly extract relevant information from a huge amount of data.

Both authors are strongly in favor of incorporating biological models of vision and learning into the field of (semi) autonomous robotics.

References

1. E. I. Barakova and T. Lourens. Prediction of rapidly changing environmental dynamics for real time behavior adaptation using visual information. Accepted for 4th Workshop on Dynamic Perception, 2002.
2. J. Bruce, T. Balch, and M. Veloso. Fast and inexpensive color image segmentation for interactive robots. In *IEEE/ISJ International Conference on Intelligent Robots and Systems (IROS 2000)*, October 2000.
3. M. Jamzad, B. S. Sadjad, V. S. Mirrokni, M. Kazemi, H. Chitsaz, A. Heydarnoori, M. T. Hajiaghai, and E. Chiniforooshan. A fast vision system for middle size robots in robocup. In *RoboCup 2001*, 2001.
4. H. Kitano, Y. Kuniyoshi, I. Noda, M. Asada, H. Matsubara, and E. Osawa. Robocup: A challenge problem for ai. *AI Magazine*, 18(1):73–85, 1997.
5. S. M. Smith and J. M. Brady. SUSAN - a new approach to low level image processing. *Int. Journal of Computer Vision*, 23(1):45–78, May 1997.

Gestalt laws and statistics

Statistics Predicts Illusions

Cornelia Fermüller and Yiannis Aloimonos

Computer Vision Laboratory
Center for Automation Research
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-3275
email : fer,yiannis@cfar.umd.edu

Abstract. The interpretation of image patterns is preceded by the detection and localization of local image features, such as line elements, intersections of line elements, and image motion. Noise in the image formation and processing, however, causes a serious problem for the estimation of features; in particular, it causes bias. As a result the location of features often is estimated erroneously. The amount of bias depends on the texture, for certain patterns it is strongly pronounced. This provides an explanation for many well-known geometrical optical illusions, such as the café wall, the Zöllner, the Poggendorff illusion and other recently discovered illusions of movement.

1 Introduction

We have found a general principle in the statistics of visual processes. Visual computations are formulated as estimation processes. Because of noise – which always is present, but very difficult to estimate accurately since visual processes involve many unknowns – these estimation processes are biased, and thus the parameters to be estimated are obtained with errors. Here we address low level estimation processes, that is edge detection, feature extraction and optical flow estimation. We argue that the bias in these low level processes is a major cause for most geometrical optical illusions.

In the past, a number of authors have discussed uncertainty in image measurements. In early studies eye movements have been advanced as a causative factor [8] in illusions. Our theory also proposes that eye movements do play a major role because they are a relevant source of noise. More recently [1, 3–5] optical or neural blur has been discussed as a cause of illusions and models of band-pass filtering or smoothing have been proposed to account for a small set of illusions [6]. In intuitive terms these studies invoked the concept explained here. Band-pass filtering constitutes a model of edge detection in noisy gray-level images. The theme of this paper is that smoothing is a special case of a more general principle – namely, uncertainty or noise causes bias in the estimation of image features – and this principle accounts for a large number of illusions that previously have been considered unrelated.

2 Errors in intensity values

Consider viewing a static scene such as the pattern in Figure 2. Let the irradiance signal coming from the scene parameterized by image position (x, y) be $I(x, y)$. The image received on the retina can be thought of as a noisy version of the ideal signal. Consider noise in the spatial location which has a Gaussian probability distribution. The expected value of the image then is obtained by smoothing, that is convolving the ideal signal with a Gaussian kernel $g(x, y, \sigma_p)$ with σ_p the standard deviation of the positional noise, that is the intensity at an image point amounts to $I(x, y) \star g(x, y, \sigma_p)$.

Edge detection mathematically amounts to localizing the extrema of the first-order derivatives or the zero crossings of second-order derivatives (the Laplacian) of the image intensity function. The change of location of straight edges under smoothing is illustrated in Figure 1. There are three cases to be considered: Edges between a dark and a bright region do not change location under scale space smoothing (Figure 1a). The two edges at the boundaries of a bright line, or bar, in a dark region (or, equivalently, a dark line in a bright region) drift apart (Figure 1b). Finally, the two edges of a line of medium brightness next to a bright and a dark region move toward each other.

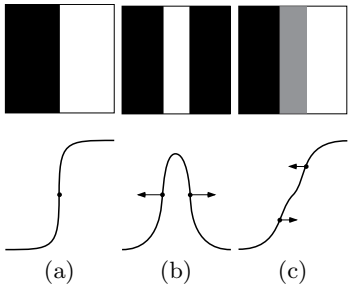
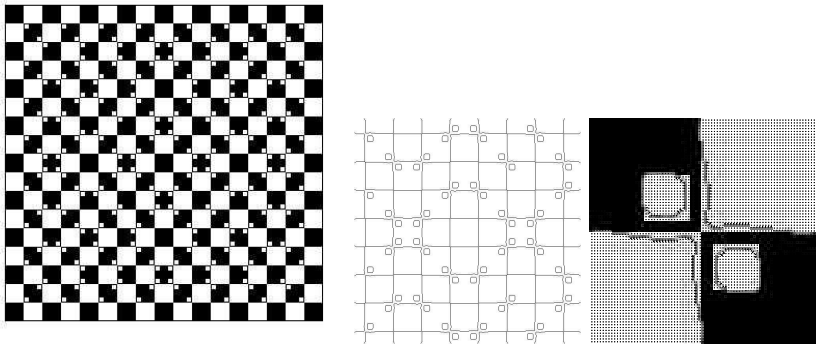


Fig. 1. A schematic description of the behavior of edge movement when smoothing: (a) no movement, (b) drifting apart, (c) getting closer.

These observations suffice to explain a number of illusions, for example the one in Figure 2a. In this pattern next to the white squares in the corners of the black squares short bars are created. The edges of these bars drift apart under smoothing and the other edges—between the black and white tiles of the checkerboard—stay in place. The result is that the edges near the locations of the white squares are bumped outward toward the white checkerboard tiles as is illustrated in Figure 2b.

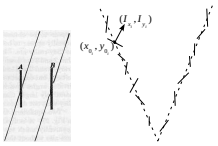
3 Errors in line elements

The perceptual effect at intersecting lines is illustrated in Figure 3a. To understand the behavior in more detail, consider the input to be edge elements.



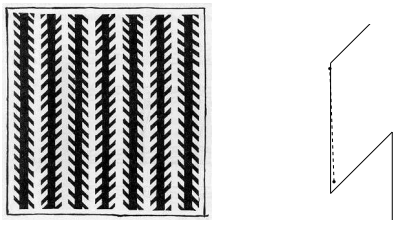
(a) (b) (c)

Fig. 2. (a) Illusory pattern: "waves." (b) The result of smoothing and edge detection on a part of the pattern. (c) The instantaneous velocity of edge points in the smoothed image – the so called drift velocity.



(a) (b)

Fig. 3. (a)From [8]. The fine line as shown in A appears to be bent in the vicinity of the broader black line, as indicated in exaggeration in B. (b) The data in the model are edgels parameterized by their center (x_{0_i}, y_{0_i}) and their direction (the unitized image gradient) (I_{x_i}, I_{y_i}).



(a) (b)

Fig. 4. (a) Zöllner pattern. (b) Estimation of edges in Zöllner pattern. The line elements are found by connecting two consecutive intersection points, resulting from the intersection of edges of two consecutive bars with the edge of the vertical bar (one in an obtuse and one in an acute angle).

A straight line is represented by a large number of edge elements (Figure 3b). These are noisy; of importance is noise in the direction. The intersection point is found by intersecting all the straight lines passing through the edge elements.

Consider additive, independently identically distributed (i.i.d.) zero-mean noise in the parameters. Let unprimed letters denote estimates and primed letters denote actual values. Each measurement i provides one equation

$$I_{x_i} x + I_{y_i} y = I_{x_i} x_{0_i} + I_{y_i} y_{0_i} \tag{1}$$

and from n measurements we obtain a system of equations which are represented in matrix form as, $I_s \boldsymbol{x} = \boldsymbol{C}$, where I_s is the n-by-2 matrix which incorporates the data in the I_{x_i} and I_{y_i}, and $\boldsymbol{C}$ is the n-dimensional vector with components $I_{x_i} x_{0_i} + I_{y_i} y_{0_i}$. The vector $\boldsymbol{x}$ denotes the intersection point whose components are x and y. The solution to the intersection point using standard least square (LS) estimation is given by

$$\boldsymbol{x} = (I_s^t I_s)^{-1} I_s^t \boldsymbol{C} \tag{2}$$

It is well known [2] that the LS solution to a linear system with errors in the measurement matrix is biased. The expected value of $\boldsymbol{x}$ is found by developing (2) into a second-order Taylor expansion at zero noise. It converges in probability to

$$\boldsymbol{x} = \boldsymbol{x}' + n {M'}^{-1} (\bar{\boldsymbol{x}}_0 - \boldsymbol{x}') \sigma_s^2 \tag{3}$$

where $M' = {I_s}'^t {I_s}'$, $\boldsymbol{x}'$ is the actual intersection point, $\bar{\boldsymbol{x}}_0$ is the mean of the $\boldsymbol{x}_{0_i}$, and σ_s^2 is the variance of the noise in the spatial derivatives of I. This expression allows for an interpretation of the bias and it allows to predict parametric influences on the strength of illusions. Some important characteristic features of the intersection of two straight lines in an acute angle are: as shown before in Figure 3 the estimated intersection is between the lines, the size of the bias decreases as the angle increases and the component of the bias in the direction perpendicular to a line decreases as the number of edgels along the line increases.

Figure 4a shows a version of the Zöllner illusion. The vertical bands are all parallel, but they look convergent or divergent. The biases in the intersection points of the edges of the bands with the edges of the short line segments cause the edge elements along the long edges between intersection points to be tilted, as illustrated in Figure 4b. A full account of the perception of tilted lines requires also an explanation of the linking of the local elements into longer lines. Our hypothesis is that this integration is computationally an approximation of the longer lines using as input the positions and orientations of the short line elements; this will give rise to tilted lines.

The model also predicts the findings of parametric studies that the illusory effect decreases with an increase in the acute angle between the main line and the obliques and that the illusion is stronger when rotated by 45 degrees, because it has been found that there is more response from the cortex to lines in horizontal and vertical than oblique orientations —translated to our model, more response means more edgels.

4 Errors in Motion

The basic image representation of movement is the optical flow which is derived in a two-stage process. First, from local spatio-temporal measurements at a point the velocity component at a point perpendicular to linear features (the normal flow) is computed. Second, normal flow measurements from features in different directions within a small neighborhood are combined to estimate the optical flow, but this estimate is biased.

We consider a gradient-based approach to deriving the normal flow. Denoting the derivatives of the image gray level $I(x, y, t)$ by I_x, I_y, I_t, and the optical flow of an image point in the x- and y-directions by $\boldsymbol{u} = (u, v)$, the following constraint is obtained:

$$I_x u + I_y v + I_t = 0 \tag{4}$$

We assume the optical flow to be constant within a region and thus obtain an over-determined system of equations whose least-squares solution amounts to

$$\boldsymbol{u} = -(I_s^t I_s)^{-1} I_s^t \boldsymbol{I}_t. \tag{5}$$

The expected value of the flow converges to

$$\boldsymbol{u} = \boldsymbol{u}' - n\sigma_s^2 M'^{-1} \boldsymbol{u}'. \tag{6}$$

Equation (6) shows the bias depends on the gradient distribution (that is, the texture) in the region with the flow always underestimated in length.

Figure 4a shows a variant of a pattern created by the graphics artist Ouchi. It consists of two rectangular checkerboard patterns oriented in orthogonal directions – a background orientation surrounding an inner ring. Small retinal motions, or slight movements of the paper, cause a segmentation of the inset pattern, and motion of the inset relative to the surround.

The tiles used to make up the pattern are longer than they are wide leading to a gradient distribution in a small region with many more normal flow measurements in one direction than the other. Since the tiles in the two regions of the figure have different orientations, the estimated regional optical flow vectors are different. The difference between the bias in the inset and the bias in the surrounding is interpreted as motion of the ring.

Another impressive illusory pattern is shown in Figure 6 (from [7]). If fixating on the center and moving the page back and forth along the line of sight the inner circle appears to rotate. This can be accounted for by different biases in the inner and outer ring; the difference in the motion vectors is tangential to the circles giving rise to the perception of a rotational motion.

5 Concluding remarks

In this paper we have discussed a major hurdle that vision systems have to deal with. Noise in the image data—that is, the image gray level and its derivatives—causes a serious problem for early visual processes and unavoidably leads to bias.

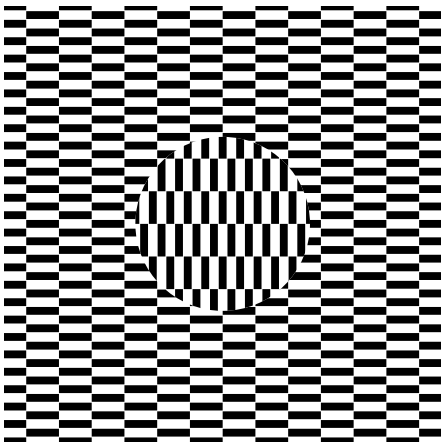


Fig. 5. A slight jiggling of the paper produces two motions.

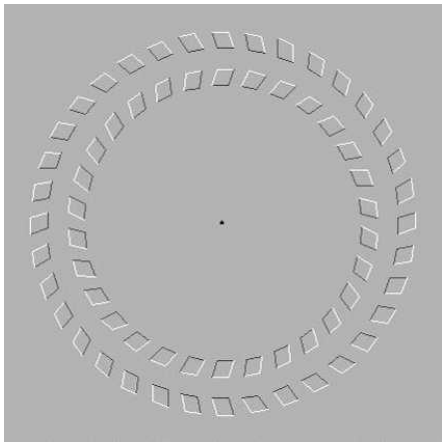


Fig. 6. Fixation at the center and movement of the figure along the line of sight causes the inner circle to rotate.

An artifact of the bias is illusory perceptions involving patterns where the bias is highly pronounced. Noise is present in any visual data. It is due to the sensing process, and in particular the spatial and temporal integration of data that moving systems are confronted with, and due to the operations involved in computing derivatives, or in estimating and locating certain frequency components of the signal. The problem is that the noise parameters usually cannot be estimated well, as they change with the lighting and viewing conditions, often too rapidly to allow enough data to be collected.

References

1. C. Chiang. A new theory to explain geometrical illusions produced by crossing lines. *Percept. Psychophys.*, 3:174–176, 1968.
2. W. Fuller. *Measurement Error Models.* Wiley, New York, 1987.
3. A. P. Ginsburg. Is the illusory triangle physical or imaginary? *Nature*, 257:219–220, 1975.
4. L. Glass. Effect of blurring on perception of a simple geometric pattern. *Nature*, 228:1341–1342, 1970.
5. S. Grossberg and E. Mingolla. Neural dynamics of perceptual grouping: Textures, boundaries and emergent segmentations. *Perception and Psychophysics*, 38(2):141–171, 1985.
6. M. J. Morgan and B. Moulden. The Münsterberg figure and twisted cords. *Vision Research*, 26(11):1793–1800, 1986.
7. B. Pinna and G. J. Brelstaff. A new visual illusion of relative motion. *Vision Research*, 40(16):2091–2096, 2000.
8. H. L. F. von Helmholtz. *Handbuch der Physiologischen Optik.* Leopold Voss, Hamburg und Leipzig, 1896.

An analysis of the motion signal distributions generated by locomotion in a natural environment

Johannes M. Zanker [1,2] and Jochen Zeil [1]

[1] Centre for Visual Sciences, RSBS, The Australian National University, Canberra, ACT 2601, Australia
j.zanker@rhul.ac.uk & zeil@rsbs.anu.edu.au
[2] Department of Psychology, Royal Holloway, University of London, Egham, Surrey TW20 0EX, England

Abstract. Many theoretical, psychophysical, and physiological studies have addressed the question of how the optic flowfields that are generated on the retina of a moving observer can be used to control behaviour. However, most of these studies were restricted to controlled laboratory conditions, and little is known about the flowfield structure under the natural conditions, organisms operate in. We investigated the information content of natural optic flowfields, by moving a panoramic imaging device outdoors on accurately defined paths and by simulating a biologically inspired motion detector network to analyse the distribution of motion signals. We demonstrate here that the motion signals obtained under natural conditions are sparsely distributed in space and that the information on the direction of local flow vectors can be ambiguous and noisy. Spatial or temporal integration is needed to retrieve reliable information on the local motion vectors. Variations of the motion detector parameters have no major effect on the overall structure of the motion signal maps. Our approach may help to assess the environmental and computational constraints in optic flow processing.

1 Introduction

A moving observer generates a large-scale pattern of image motion on the retina that contains information on both observer movement – egomotion - and the three-dimensional structure of the environment. Gibson [1] recognised the significance of such optic flowfields, which he illustrated by arrays of homogenously distributed velocity vectors, and thus sparked the development of flowfield theory, which deals with algorithms to extract egomotion parameters from optic flow [e.g., 2, 3]. Most algorithms assume implicitly that local motion signals are veridical, homogenously distributed, and carry true velocity information. However, the actual structure of two-dimensional motion signal distributions is determined (i) by the pattern of locomotion, (ii) by the specific three-dimensional layout of the local environment, and (iii) by the motion detection mechanism employed. Similarly, motion sensitive, optic flow processing neurones in invertebrates and vertebrates [e.g., 4, 5] are usually investigated with coherently structured motion stimuli that densely cover large parts

of the visual field. In simulations, Dahmen et al. [6] have recently shown that surprisingly few and low fidelity flow measurements are needed to estimate egomotion parameters, as long as these local measurements are widely distributed throughout the visual field. To appreciate the design of neural mechanisms underlying flowfield processing and to assess the robustness of optic low algorithms, we thus need to know in more detail, what kind of motion signal distributions visual systems are confronted with in a normal ecological and behavioural context.

To address this issue, we studied the role of environmental and computational constraints in natural optic flow processing by moving a panoramic imaging device along precisely defined three-dimensional paths in a variety of natural outdoor locations. The recorded image sequences then served as input to a biologically inspired, two-dimensional motion detector network (2DMD), consisting of an array of correlation-type detector pairs for horizontal and vertical motion components [7]. This procedure generates panoramic motion signal maps which allow us to study the structure and dynamics of natural motion signal distributions, as they would be experienced 'in the cockpit' of a low-flying observer, like an insect.

2 Methods

A panoramic imaging device, mounted on a computer-controlled robotic gantry, was moved by means of DC servo-motors along accurately defined 3D-trajectories within a space of about 1 m^3 with a positioning accuracy of 0.1 mm^3. The imaging device consisted of a black and white video camera (Samsung BW-410CA) looking down onto a parabolic mirror which was optimised for constant spatial resolution [8]. Images were digitised (8 bit, Matrox Meteor framegrabber) and stored on a computer for off-line analysis. The raw images, in which azimuth ϕ and elevation θ are represented in polar coordinates (see figure 1A), were converted into Cartesian coordinates (unwarping software by courtesy of Javaan Chahl), leading to images 450 pixels wide and 185 pixels high (corresponding to a visual field size of $\phi = 360°$ and $\theta = 136°$, figure 1B). In the default configuration, image sequences of 64 consecutive frames were taken at 25 frames/s during gantry speeds of 5 cm/s and 10 cm/s.

Image sequences were analysed with a two-dimensional motion detector model (2DMD) which has previously been used to simulate a variety of behavioural and psychophysical phenomena [e.g., 7, 9, 10]. The basic building blocks of the 2DMD model are elementary motion detectors (EMDs) of the correlation type which have been shown to be good candidates for biologically implemented motion detectors [for review, see 11] and are representative of a variety of luminance based motion detection algorithms [e.g., 12]. In a simple implementation (figure 1C), each EMD receives input from two points of the spatially filtered stimulus patterns. To remove DC components from the input, difference of Gaussians (DOGs) with balanced excitatory centre and inhibitory surround are used as bandpass filters in the input lines. The fundamental spatial model parameter was the sampling distance $\Delta\varphi$ between the two inputs (2 pixels, approximately 1.6°, as default). To prevent aliasing, the diameter of the receptive field was set to about twice the value of the sampling distance. The signal from one input line is multiplied with the temporally filtered

signal from the other line, and the outputs of two antisymmetric units of this kind are subtracted from each other with equal weights, leading to a fully opponent, and thus highly directionally selective EMD. The fundamental temporal model parameter was the time constant τ of the first-order lowpass filter (2 frame intervals, 80 ms, as default). An increased temporal resolution was used in the simulations (the frame interval corresponding to 8 digital simulation steps) to improve the accuracy in calculating the dynamic responses of the temporal filters.

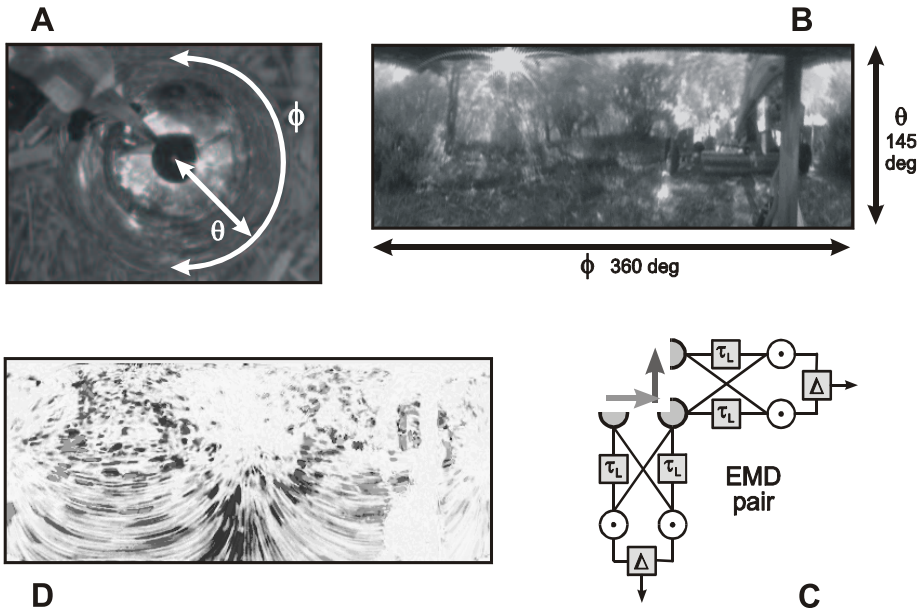


Figure 1: A method to study natural optic flowfields. A video camera is used to capture panoramic images in polar coordinates (**A**) which are converted into Cartesian coordinates (**B**; azimuth Φ, elevation θ). Image sequences are recorded while moving the camera through a natural scene and then used as input to a large array of motion detector pairs (one element sketched in **C**), to generate motion signal maps (**D**).

Movie sequences were processed by two arrays of such EMDs, which were oriented along the horizontal and vertical Cartesian image axis, respectively (sketched in fig. 1C) The 2DMD model thus consists of two sets of 450 x 185 correlators, one pair centred at each image pixel. The output of the model is a two-dimensional motion signal distribution, which we call a *motion signal map*, with a horizontal and vertical signal component for each image point (see fig. 1D). In some cases this raw 2DMD output was temporally averaged (over 8 to 64 frames) before further analysis. We use a two-dimensional colour code to represent the direction and the magnitude of local motion detector responses in these motion signal maps in terms of hue and saturation [10].

3 Results and Discussion

The 2DMD response for a simple forward translation is compared in figure 2 for a single displacement step (A) to the average of 16 consecutive steps (B) and for a variety of EMD parameter settings (C-F). The motion signal maps are characterised by a systematic pattern of colour change (i.e. a change in the direction of local image motion) around the centre of each panel, reflecting the local motion vectors radiating from the flowfield pole. Local image motion directions are inverted in the image regions corresponding to the rear field of view, close to the left and right borders of the panoramic image. This distribution of local motion signals is typical for a forward translation of the camera which produces an expanding and contracting flowfield pole in the frontal and the rear field of view, respectively. Corresponding patterns of local motion signals, with different locations of the flowfield pole in the images, are found for translations in other directions (data not shown).

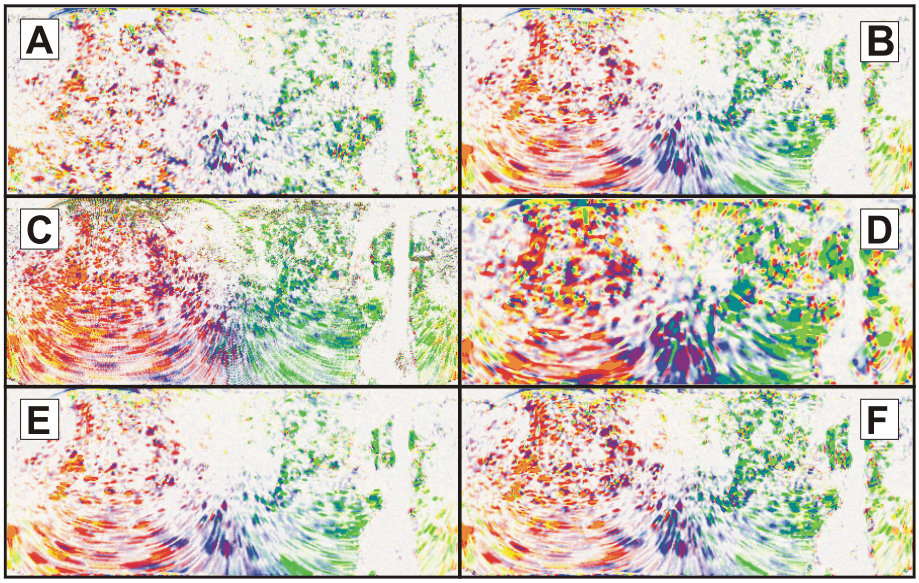


Figure 2: Motion signal maps derived as 2DMD output (**A**: single frame **B-F**: average of 16 consecutive frames) for a variety of EMD model parameters. **A-B**: $\Delta\varphi = 2$ pixel, $\tau = 2$ frames; **C**: $\Delta\varphi = 1$, $\tau = 2$; **D**: $\Delta\varphi = 4$, $\tau = 2$; **E**: $\Delta\varphi = 2$, $\tau = 1$; **F**: $\Delta\varphi = 2$, $\tau = 4$. Each panel shows the output of a set of 450 x 185 EMD pairs (360° azimuth, 136° elevation) in 2D-colour code representing direction and strength of the local motion signal (green-right, yellow-up, red-left, blue-down).

Although the overall structure of the flowfield can be recognised in the individual 2DMD output frame, it is striking how noisy and sparse the distribution of local motion signals is in such cluttered natural scenes (fig. 2A). Most importantly, the image regions around the flowfield poles do not contain clear motion signals, which can be explained by considering the fact that image speed is minimal there, thus attenuating the local EMD response which depends on the contrast and speed of

moving contours. The directional noise apparent in these motion signal maps be attributed to fluctuations inherent to the EMD output [13] and to variations of local contour orientations which affect the detected direction of motion [14]. The flowfield structure in motion signal maps can be improved by spatial or temporal averaging (compare fig. 2A and B), reducing some of the noise in local motion signals, and their sparseness, at the cost of resolution. We find in addition that the density of the maps is affected by the gain and the non-linearity of amplifying the EMD output (data not shown). Since spatial and temporal EMD parameters are known to determine the tuning of the detector to spatial frquency and image velocity [e.g., 15], we investigated how variations of the time constant τ and the sampling distance $\Delta\varphi$ affect the structure of the motion signal maps. The examples shown in figure 2 C-F demonstrate that $\Delta\varphi$ has some (obvious) influence on the spatial grain and noise load of the motion signal maps (cf. fig. 2C and D), but very little effect on the overall structure. The effects of changing τ are negligible (cf. fig. 2E and F). This resistance against variations in model parameter can be related to the inherently broadband properties of objects and surfaces in natural scenes [16].

Natural motion signal maps have two additional properties which are relevant for otpic flow processing. Firstly, contours on the ground generate comparatively large image motion components compared with more distant objects above the horizon. The 'ventral' regions of the motion signal maps are thus occupied by denser and stronger motion signals than the 'dorsal' regions. Secondly, when several 2DMD response frames are averaged (fig. 2B-E), the motion signals are aligned along 'motion streaks', which reflect the trajectories of image contrast elements during the averaging interval. These oriented streaks contain independent information on the structure of optic flow - the radial patterns in the front and the rear image regions provide a clear indication of the flowfield poles. Recent psychophysical experiments suggest that humans are actually able to use the orientation of temporally blurred moving objects for motion processing [17].

4 Conclusions

The three-dimensional layout of the environment, as defined by the size, the texture, the contrast, the density, and the distance of objects, has profound effects on the local motion signals in the visual field of a moving observer [6]. Mechanisms to extract relevant information from natural optic flow fields are likely to be adapted to the lifestyle, and in particular to the locomotion patterns of animals, as well as to the statistical properties of the world they inhabit [18]. Our results indicate that in natural scenes the motion signals generated by translational movements are sparse and noisy, but that egomotion parameters can be estimated at a coarse spatial or temporal scale from the radiating pattern of local motion directions. Interestingly, the overall structure of the motion signal distribution is rather robust against variations of the basic EMD model parameters. Future work has to show how well these results generalise to other environments and more complex types of locomotion. In the context of both biological and machine vision, our approach to generate dynamic motion signal maps under realistic operating conditions can critically extend earlier

attempts [e.g., 19] to test the reliability and robustness of techniques, algorithms and computational models of optic flow processing.

References:

1. Gibson, J. J., The perception of the visual world. The Riverside Press, Cambridge, MA (1950)
2. Koenderink, J. J.Van Doorn, A. J.: Facts on Optic Flow. Biol.Cybern. 56 (1987) 247-254
3. Longuet-Higgins, H. C.Prazdny, K.: The interpretation of a moving retinal image. Proc.R.Soc.Lond B 208 (1980) 385-397
4. Hausen, K.Egelhaaf, M.: Neural Mechanisms of Visual Course Control in Insects. In: D. G. Stavenga, R. C. Hardie (eds): Facets of Vision. Springer Verlag, Berlin Heidelberg (1989) 391-424
5. Frost, B. J., Wylie, D. R., Wang, Y.-C.: The processing of object and self-motion in the tectofugal and accesory optic pathways of birds. Vision Res. 30 (1990) 1677-1688
6. Dahmen, H. J., Franz, M. O., Krapp, H. G.: Extracting egomotion from optic flow: limits of accuracy and neural matched filters. In: J. M. Zanker, J. Zeil (eds): Motion Vision - Computational, Neural, and Ecological Constraints. Springer, Berlin Heidelberg New York (2001) 143-168
7. Zanker, J. M., Hofmann, M. I., Zeil, J.: A two-dimensional motion detector model (2DMD) responding to artificial and natural image sequences. Invest.Ophth.Vis.Science 38 (1997) S 936
8. Chahl, J. S.Srinivasan, M. V.: Reflective surfaces for panoramic imaging. Appl.Opt. 36 (1997) 8275-8285
9. Zanker, J. M.: Combining Local Motion Signals: A Computational Study of Segmentation and Transparency. In: J. M. Zanker, J. Zeil (eds): Motion Vision: Computational, Neural and Ecological Constraints. Springer, Berlin Heidelberg New York (2001)
10. Zeil, J.Zanker, J. M.: A Glimpse into Crabworld. Vision Research 37 (1997) 3417-3426
11. Borst, A.Egelhaaf, M.: Principles of visual motion detection. Trends in Neuroscience 12 (1989) 297-306
12. Adelson, E. H.Bergen, J. R.: Spatiotemporal energy models for the perception of motion. J.Opt.Soc.Am. A 2 (1985) 284-299
13. Reichardt, W.Egelhaaf, M.: Properties of Individual Movement Detectors as Derived from Behavioural Experiments on the Visual System of the Fly. Biol.Cybern. 58 (1988) 287-294
14. Hildreth, E.-C.Koch, C.: The analysis of visual motion: From computational theory to neuronal mechanisms. Ann.Rev.Neurosci. 10 (1987) 477-533
15. Zanker, J. M., Srinivasan, M. V., Egelhaaf, M.: Speed tuning in elementary motion detectors of the correlation type. Biol.Cybern. 80 (1999) 109-116
16. Field, D. J.: Relations between the statistics of natural images and the response properties of cortical cells. J.Opt.Soc.Am. A4 (1987) 2379-2394
17. Geisler, W. S.: Motion streaks provide a spatial code for motion direction. Nature 400 (1999) 65-69
18. Eckert, M. P.Zeil, J.: Towards an ecology of motion vision. In: J. M. Zanker, J. Zeil (eds): Motion Vision: Computational, neural and ecological constraints. Springer Verlag, Berlin Heidelberg New York (2001) 333-369
19. Barron, J. L., Fleet, D. J., Beauchemin, S. S.: Performance of Optical Flow Techniques. International Journal of Computer Vision 12 (1994) 43-77

Stimulus Sensitivity in Monkey Visual Cortex is Modulated by Viewing Distance while Spatial Frequency Tuning and Receptive Field Size are not

Hans Jörg Brinksmeyer, Frank Michler, Alexander Gail, and Reinhard Eckhorn

Group of NeuroPhysics, Department of Physics, Philipps-University, Renthof 7, D-35032 Marburg, Germany
corresponding author: reinhard.eckhorn@physik.uni-marburg.de

Abstract. We searched for neural mechanisms allowing for distance invariant object processing in visual cortex. Such mechanisms may require modulation of response properties in visual cortical neurons with viewing distance, including response sensitivity, size of the classical receptive field (**CRF**), and preference for spatial frequency (**SF**). In order to test these hypotheses we recorded multiple unit activity (**MUA**) in primary and secondary visual cortex (V1, V2) of an awake macaque monkey with an array of 16 microelectrodes while changing randomly the viewing distance by moving the stimulus monitor. We confirmed previous work reporting strong sensitivity modulation with changes in viewing distance of *near*-, *intermediate*-, and *far*-type [1, 2]. In contrast, we found CRF-size and SF-preference on average being independent of viewing distance. This suggests that distance invariant object coding is supported by a subset of V1 and V2 neurons, probably selected by facilitation via neuronal input representing a distance estimate. Our data further suggests that neurons with overlapping CRFs and different preferred SFs mutually couple their activities in order to code for the spatial profiles of local luminance contrast at object contours. We found shortest response delays to low and medium SFs while those to high SFs were significantly longer.

1 Introduction

We are interested in the neural mechanisms of visual size invariance. Under natural viewing conditions size invariance requires a mechanism allowing for distance-invariant recognition of objects. The term distance-invariance refers to toleration of changes in retinal image size that are due to varying viewing distance as opposed to varying real-world object size (see related model in this Volume [3]). The required invariance transformation is probably learned during everyday experience. Basis for this is our knowledge of objects keeping their identity and physical size with viewing distance. Psychophysical work demonstrated that humans can estimate the size of objects rather precisely up to 30 m

distance and that this capability is correlated with the precise estimate of distance [4–6]. Thus, a neural mechanism estimating absolute object distance may tune the invariance-mechanism with this single parameter like a photographer tunes the setting of his tele-objective in order to compensate for object distance. Such tuning can be achieved by different neural mechanisms. One is proposed by Kupper & Eckhorn [3], introducing distance complex cells that receive input from sets of lower-level feature detectors, modulated in their sensitivity by distance. Such distance-dependent modulations of neurons have recently been reported for visual cortical areas V1, V2 and V4 of awake monkeys [1,2]. As visual objects appear under increasingly smaller viewing angle with increasing viewing distance, neurons involved in distance invariance operations might modulate the size of their classical receptive fields (**CRFs**) and/or their preferred spatial frequency (**SF**): tuning their CRFs to smaller size and higher spatial frequencies when fixated objects appear far, and modulating the CRFs to large size and lower SF for near viewing distance.

2 Methods

In order to test these hypotheses we recorded multiple unit activity (**MUA**) at parafoveal representations in the upper layers of striate (V1) and extrastriate (V2) visual cortex of a macaque monkey with an array of 16 microelectrodes in each session. In all tasks the monkey was rewarded for steadily fixating a small luminance spot with a maximal error of $\pm 0.5^\circ$ visual angle. Eye positions were controlled and recorded by an infrared camera system (resolutions: 0.05°, 225 Hz). For the investigation of distance dependent effects visual luminance stimuli were presented pseudo-randomly at three different distances (0.45, 0.9, 1.8 m, and 0.45, 0.78, 1.35 m, respectively) via a computer screen (TFT) quickly movable along the axis of straight sight under computer control (up to 1 m/s). For each recording site we determined at each stimulus distance (1) the positions, sizes, and sensitivities of CRFs, probed by a randomly jumping spot (RF-cinematogram method: [7]), size-scaled to the different distances; (2) the orientation- and SF-tuning, probed at random order with Gabor-patches of different orientations and SFs; (3) the signal interactions among SF-channels having their CRFs at the contour of a luminance-defined object that was radial symmetric with respect to the fixation point and was presented size-scaled at the three distances. All measures were analyzed with respect to their dependency on viewing distance.

3 Results

Distance characteristics. The response strength at most recording sites in V1 and V2 was modulated by viewing distance. Depending on the distance where the response maximum appears we classified *far*-, *intermediate*-, and *near*-distance characteristics, based on the three stimulation distances used (Fig. 1A). The others were classified *non-modulated*. Most recording locations displayed the near

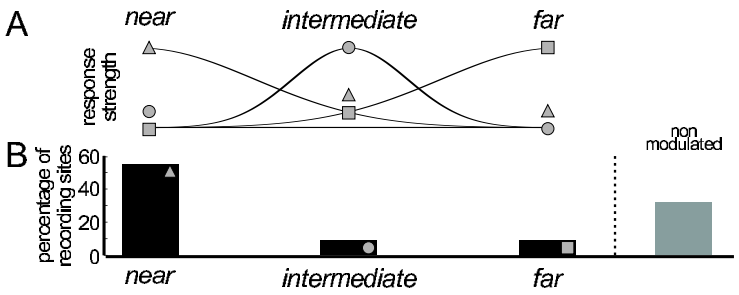


Fig. 1. A: Schematic tuning characteristics for viewing distance: *near*, *intermediate*, *far*. **B:** Percentage of recording locations for the four distance characteristics, examplary for the V2 recordings.

characteristic (56%; Fig. 1B). We found neurons of any SF-preference within each class of distance modulation.

SF-preference and CRF-size. SF-preferences cover a broad range of more than 1 to 8 (> 3 octaves; SD = 3.04 cyc/° (V1), 3.40 cyc/° (V2)) already within each of the small visual field representations measured by us in V1 (0.5–3° horiz.; 1–2° vert.) and V2 (0–1.5° horiz.; 1.7–4° vert.), respectively (Fig. 2A). However, within the same recording chamber CRF-sizes span only a range of about 1 to 2 (1 octave; SD = 0.07° (V1), 0.15° (V2)). Our data show no correlation among SF-preference and CRF-size, as might be expected (CC = −0.28 (V1), −0.13 (V2)). In addition, neither average SF-preference (Fig. 2C) nor average CRF-size (Fig. 2B) depended on viewing distance. However, the individual CRF-sizes of many recording positions showed some change with stimulus distance (21% standard deviation of the mean CRF-size).

Response latency in V1 and V2 to stimulus onset of a Gabor patch depends in our data on its SF (Fig. 3): on average stimuli at low and medium SFs caused shorter response latencies (about 70 ms; 0.5–4 cyc/°) compared to high SFs (about 92 ms; 10–13 cyc/°). Note that this dependency characterizes the average delay at a given recording location to Gabor stimuli with different SFs.

Signal coupling at object contour. With activations by a luminance contour, neurons with overlapping CRFs and with orientation preference matching that of the contour, mutually couple their signals in the frequency ranges 10–25 Hz and 40–50 Hz, shortly after the transient broad-band response (Fig. 4A). These beta- and gamma-range couplings are present independently of the SF-preferences at the recording locations, including recording pairs with similar and different SF-preferences. In contrast, recording pairs with orientation preferences orthogonal to the contour lack this type of coupling (Fig. 4B). More subtle analysis of the orthogonal class indicates decreased correlation after the transient response. Additional properties of signal couplings among neurons of defined SF-preference have not been analyzed by us yet.

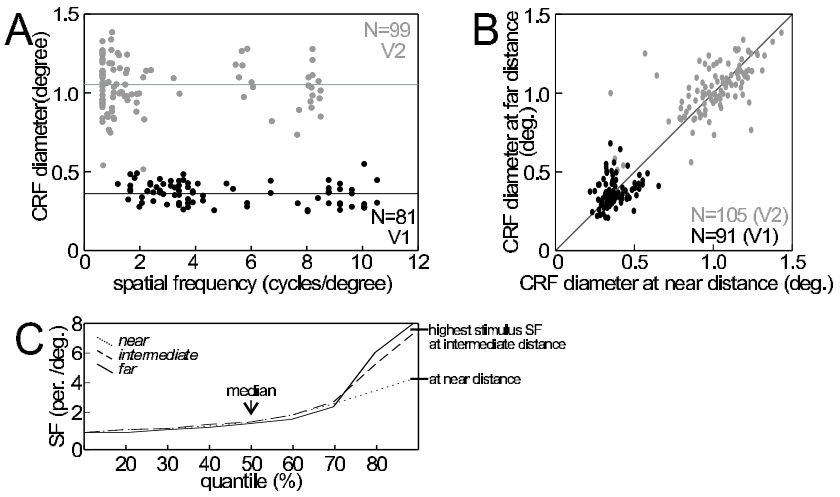


Fig. 2. A: Dependency of spatial frequency (SF) preference on the diameter of the classical receptive fields (CRF) in V1 and V2. **B:** Distribution of CRF-diameters with *near* against *far* stimulation. **C:** Distribution of SF-preference for *near*, *intermediate* and *far* stimulation. Note that the clustering of data is due to different recording areas (V1 vs. V2) and recording excentricity from the fovea.

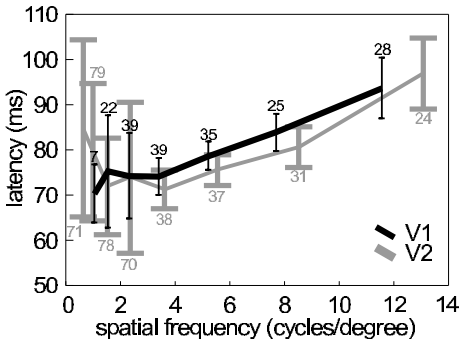


Fig. 3. Response latencies to the presentation of Gabor patches are shortest to medium and low spatial frequencies (SFs) and longest to high SFs in V1 and V2.

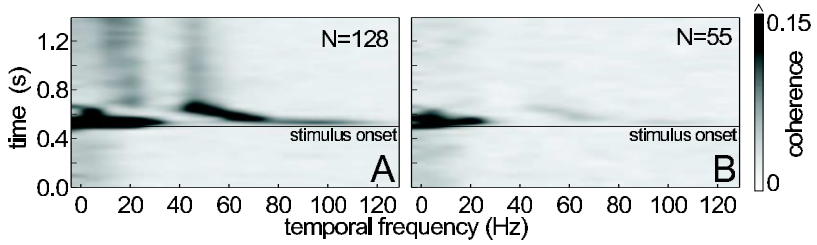


Fig. 4. Coupling dynamics among recording pairs with overlapping CRFs at a luminance contour quantified by spectral coherence (gray code). **A:** Orientation preference matches contour orientation. **B:** Orientation preference orthogonal to contour.

4 Discussion and Conclusions

Distinct types of distance tuning. Confirming recent experimental work, and in line with the suggestion of a model of distance invariance [3], we found three distinct classes of distance selectivity: *near*-, *intermediate*-, and *far*-tuned (Fig. 1A). These modulations are due to neural input changing with absolute viewing distance and may be based on a broad variety of distance cues, including ocular vergence and perspective. Vergence angle can provide an absolute cue for fixation distance (at least up to about 5 meters). We varied it systematically with stimulus distance because the monkey was always required to fixate the screen. From the same reason ocular disparity was kept constant with different distances and therefore can probably play no role in distance dependent modulations.

Only small changes in CRF-size and SF-tuning with viewing distance. Visual objects appear under increasingly smaller viewing angle with increasing viewing distance. We might therefore expect that neurons involved in distance invariance operations decrease their average CRFs' sizes with distance in order to cope with the higher resolution required for the far-appearing object. Our data do not support this expectation: average CRF size does not vary with distance. Another finding is unexpected and shows that linear filter theory is probably not appropriate for applications to cortical SF-tuning: preferred SFs within a small cortical range of recording locations and hence, within a small range of visual field representation, vary over a broad range (more than eight-fold) while the CRF-diameters at these same locations vary only by a factor of about two. Hence, SF-preference does not scale with CRF-diameter (Fig. 2A). In addition, average SF-preference at the recording locations does not change with distance. For distance invariant object coding, modulations of SF-preference and CRF-size with distance may not be required for a given location of visual excentricity (where we performed our recordings) because the contour of a fixated object changes its representation in the visual system automatically from lower to higher resolution when the object size shrinks with distance and the contour activates the high-resolution central neurons. Our results on CRF-size and SF-tuning on distance are with high probability not an effect of MUA- (instead of single-unit-)

recording, because in striate cortex CRF-sizes and SF-tuning are only slightly smaller with single-unit- compared to MUA-recordings.

Response latency was on average shorter for stimuli at low and medium SFs (about 70 ms; 1–4 cyc/°) and was significantly longer at high SFs (about 95 ms; 10–13 cyc/°). We assumme that the strongest effect on latency is due to retinal processing. At a given retinal excentricity the CRF-sizes and SF-preferences are rather constant. Magnocellular neurons have larger CRFs, prefer lower SFs, and transmit signals faster than parvocellular neurons. Thus, stimulation with varying SFs will activate different proportions of magno- and parvo-cells. Another effect on latency by SF is introduced by both magno- and parvo-cells: strong stimuli and hence, stimulation at their preferred SF, will cause large responses at short delays. In contrast, the used high SF stimuli probably have activated the neurons less strongly and therefore led to higher latencies.

Contour coding by SF-channels. We found the luminance step of a contour coded by neurons synchronizing at high frequency and having overlapping CRFs such that the profile of a current contour may be represented by the superposition of their CRFs. We have additional predictions but did not yet analyze our data in sufficient detail: (1) At positions of the object's surface, we expect facilitatory coupling among neurons with offset (non-overlapping) CRFs preferring low SFs. (2) At object surface representations we expect inhibition of neurons preferring high SFs by those preferring low SF with CRFs at the same position.

5 Acknowledgements

Typesetting by Basim Al-Shaikhli and financial support by the DFG (Ec 53/10-1 to R.E.) are greatly acknowledged.

References

1. Dobbins, A. C., Jeo, R. M. , Fiser, J. , Allman, J. M. Distance modulation of neural activity in the visual cortex. Science **281** (1998) 552–555
2. Rosenbluth, D., Allman, J. M.: The effect of gaze angle and fixation distance on the responses of neurons in V1, V2, and V4. Neuron **33** (2002) 143–149
3. Kupper, R., Eckhorn, R.: A neural network model generating invariance for visual distance. Abstract 4th Workshop Dynamic Perception 2002 in Bochum, 14.–15. November
4. Holway, A. H., Boring, E. G. Am. J. Psychol. **54** (1942) 21–37
5. Humphrey, N. K., Weiskrantz, L. Size constancy in monkeys with inferotemporal lesions. Quarterly Journal of Experimental Psychology **21** (1969) 225–238
6. Ungerleider, L. G., Ganz, L., Pribram, K. H. Size constancy in rhesus monkeys: Effects of pulvinar, prestriate, and inferotemporal lesions. Exp. Brain Res. **27** (1977) 251–269
7. Eckhorn, R. , Krause, F. , Nelson, J. I.: The RF-cinematogram. Biol. Cybern. **69** (1993) 37–55

Brightness Perception and Real World Image Processing - A Unifying Account

*** Matthias S. Keil[1,2†], Gabriel Cristóbal[1], and Heiko Neumann[2]

[1] Instituto de Óptica (CSIC), Image & Vision Department, E-28006 Madrid (Spain)
mat@optica.csic.es, gabriel@optica.csic.es
[2] Universität Ulm, Abteilung Neuroinformatik, Fakultät für Informatik
Albert-Einstein-Allee D-89069 Ulm (Germany)
hneumann@neuro.informatik.uni-ulm.de

Abstract. A novel single-scale neural architecture is proposed which both reproduces brightness illusions and successfully deals with natural images. Our architecture builds upon the premise that early vision should facilitate object recognition. Specifically, the visual input is segregated into three categories, namely texture (small-scale even symmetric features), surfaces (small-scale odd symmetric features) and gradients (large-scale even and odd symmetric features). The model also proposes a solution to anchoring brightness by means of a novel multiplexed retinal code. In this way a single-scale architecture is sufficient to recover absolute luminance levels.

1 Introduction

Our proposed architecture for brightness processing aims to unify two seemingly diverging goals, that is image processing and brightness perception. A successful unification has not been achieved so far, since models which predict brightness phenomena only rarely produce meaningful results when processing real-world images (although some results have been demonstrated, e.g. [1]). On the other hand, models for image processing tasks (typically coding or denoising), which often claim to provide some account to early vision, fail to predict phenomena associated with brightness perception. Usually, both model classes compute their output by superimposing processed filter outputs over various scales and orientations, whereby filter outputs are processed in order to fulfill a certain pre-defined goal (coding, denoising, predicting psychophysical results, etc.). None of these models has achieved any segregation of the visual input in way compatible with object recognition; rather, these models create only an internal (or cortical) representation of the visual input, thus deferring segregation to higher level cortical processing.

Furthermore, there is no model available for processing two-dimensional luminance patterns which comes up with a neurophysiological plausible solution to

*** This research is supported by the German-Spanish Academic Research Collaboration Program (DAAD, acciones integradas Hispano- Alemanes 2002/03, Proyecto No.HA2001-0087

† Corresponding author

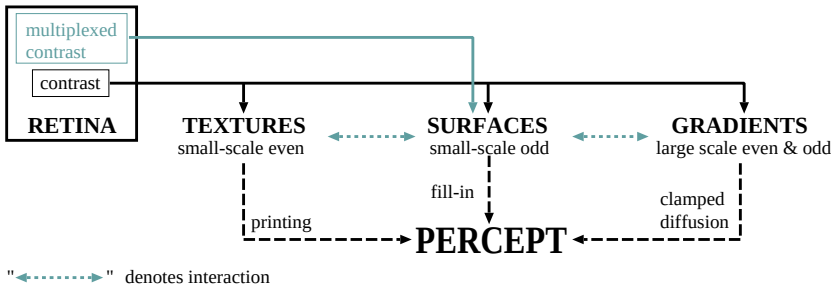


Fig. 1. Sketch of the proposed architecture. Dotted lines denote stages which were not implemented yet. Specifically, suitable interactions between the three subsystems may be defined to improve the segregation process.

the *anchoring problem* (although a one-dimensional solution was suggested by [2]). This problem is commonly solved by employing an additional "luminance channel" in the form of a low-passed filtered (or large-scale band-passed, e.g. [1]) version of the visual input, e.g. [3–7]. Yet, evidence supporting the existence of such a channel is still lacking.

Here we present a novel architecture for foveal brightness perception in agreement with known neurophysiological data (see figure 1). We propose that cortical simple cells of different symmetries (even, odd) and scales extract different aspects from the visual input, which are *(i) texture* (here defined as small-scale even symmetric features, such as lines and points), *(ii) surfaces* (corresponding to small-scale odd symmetric features for building cortical surface representations), and *(iii)* luminance *gradients* (corresponding to large-scale even and odd symmetric features, for example out-of-focus lines or edges). Simulations show how this segregation process renders cortical representations of object surfaces invariant to noise and illumination gradients.

Also, we suggest a neurophysiologically plausible solution to the anchoring problem by proposing a "multiplexed" retinal code which at the same time represents information about contrast and brightness (ON-cell) and contrast and darkness (OFF-cell) of a visual input.

2 A new model for brightness perception

Our architecture builds upon filling-in theory [8]. It consists of a retinal stage, and three cortical stages. Each cortical stage consists of two layers, where activity in one layer is thought to correspond to brightness, and activity in the other layer is thought to correspond to darkness. The perceptual activity (or perceived luminance) is essentially computed by subtracting darkness from, and adding brightness to, an Eigengrau value [9, 10]. A brief description of the individual model stages is given below.

Retinal Processing. [11, 12] found evidence that in addition to the center/surround receptive field of retinal ganglion cells there exists a disinhibitory region or *outer surround*. This region corresponds to an annulus

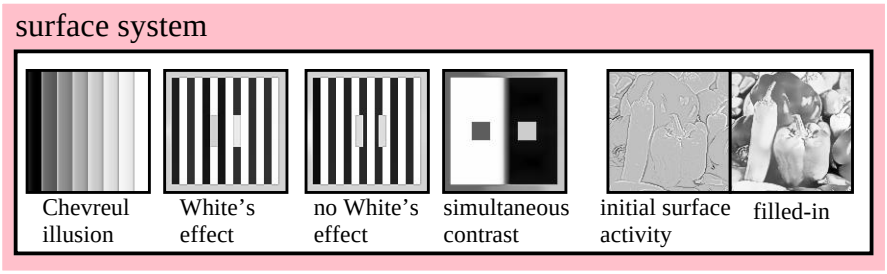


Fig. 2. Results for the surface system. Left: Simulation results for brightness illusions. **Right:** The filled-in result of a real-world image is juxtaposed with gated multiplexed activity which corresponds to the initial state of the brightness map (denoted by brighter colors) and the darkness map (darker colors).

around a ganglion cell's center/surround receptive field. In our model we employ an outer surround as a measure of local luminance, which is used to modulate response amplitudes of retinal ganglion cells such that an ON-cell (OFF-cell) contains both information about contrast and local brightness (darkness). In this way a multiplexed retinal code is created. Notice that in this way we are able to convey information about absolute luminance levels with one single scale (in fact, we model only foveal vision, that is we use the smallest possible receptive fields), whereas usually large filter scales are employed for this purpose (the center corresponds to the visual input, the surround to its four nearest neighbors, and the outer surround to a Gaussian with a spatial constant $\sigma = 4$ pixels). The multiplexed retinal code provides a solution to the anchoring problem.

Surface system. Odd-symmetric contrast configurations in the visual input (typically edges) trigger the *gating* of multiplexed retinal activity into surface layers. Surface representations are built by means of a novel diffusion paradigm which fills-in the gated multiplexed activity in corresponding filling-in domains. Filling-in domains are defined by odd symmetric contrast borders, which eventually correspond to surface representations. Instead of heat diffusion as filling-in mechanism [8], we propose a novel diffusion paradigm which converges in shorter time to homogeneously filled-in surface representations, and which discounts large scale activity gradients. The new diffusion equations for brightness activity s°_{ij} and darkness activity $s^{\bullet}_{\mathrm{ij}}$ are given by

$$\begin{aligned} \frac{ds^{\circ}_{\mathrm{ij}}(t)}{dt} &= \gamma_w w^{\bullet}_{ij}(E_{in} - s^{\circ}_{\mathrm{ij}}) + \mathcal{K}^{\circ}_{\epsilon,\infty} s^{\circ}_{\mathrm{ij}} + \delta(t - t_0)\tilde{m}^{\oplus}_{ij} \\ \frac{ds^{\bullet}_{\mathrm{ij}}(t)}{dt} &= \gamma_w w^{\circ}_{ij}(E_{in} - s^{\bullet}_{\mathrm{ij}}) + \mathcal{K}^{\bullet}_{\epsilon,\infty} s^{\bullet}_{\mathrm{ij}} + \delta(t - t_0)\tilde{m}^{\ominus}_{ij} \end{aligned} \tag{1}$$

where γ_w is a constant synaptic weight, $w^{\bullet}_{ij}$ and w°_{ij} are two sets of odd symmetric boundaries, E_{in} is a inhibitory reversal potential (i.e. boundaries hyperpolarize the membrane potential of surface cells), $\mathcal{K}^{\circ}_{\epsilon,\infty}$ and $\mathcal{K}^{\bullet}_{\epsilon,\infty}$ are nonlinear diffusion operators (which include mutual inhibition of brightness and darkness activity), $\delta(t - t_0)$ is Dirac's delta function, and finally $\tilde{m}^{\oplus}_{ij}$

and $\tilde{m}_{ij}^{\ominus}$ are multiplexed retinal activities (filling-in of multiplexed retinal activities instantaneously recovers absolute luminance values).
Many brightness illusions (such as White's effect, grating induction, Benary cross, simultaneous brightness contrast) are reproduced by the surface system (figure 2). Specifically, the surface system provides a new account to White's effect and the Benary cross; these illusions occur as a consequence of the multiplexed retinal code and the novel diffusion paradigm.

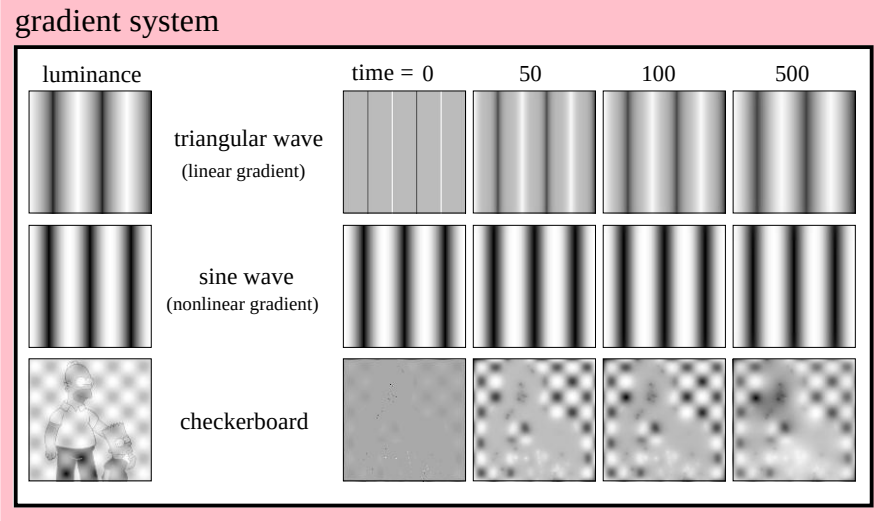


Fig. 3. Results for the gradient system. Snapshots at different time steps (see numbers) show the evolution of the perceptual activity in the gradient maps. The input ("luminance") is shown in the left image of each row. The first row shows the generation of a luminance gradient with linear slope, where Mach-like bands were generated in the output. In the second row it is shown that with a nonlinear luminance gradient (here a sine wave grating), no explicit generation of a gradient is observed, since the state of the gradient system remains approximately stationary. The example in the last row illustrates that surfaces are suppressed, but gradients are represented in the gradient system.

Gradient system. Gradients may contain valuable information about 3-D surface structure (structure from shading, e.g. [13]) and therefore provide additional information for object recognition. Gradients are defined as large-scale even and odd symmetric features. Since our model only employs a single scale, we have to recover large-scale gradients by means of *clamped diffusion* (see figure 3). This process works in a way that in the brightness layer ON-activity serves as tonic (or "clamped") source, and the OFF-activity as tonic sink (vice versa for the gradient darkness layer).
The gradient system successfully accounts for the inverted-U behavior of the perceived strength of Mach bands vs. the slope (or spatial frequency) of the luminance ramp (i.e. there exists a ramp slope where Mach bands are perceived with maximum strength) [14].

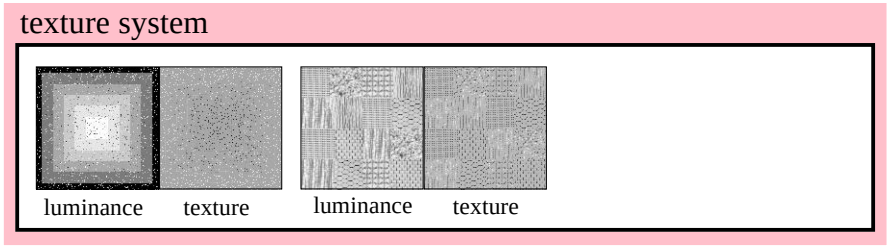


Fig. 4. Results for the texture system. Left: speckled noise was added to a luminance pyramid, where each pixel received noise with probability $p = 0.1$ ("luminance"). The output of the texture system shows a successful segregation of even-symmetric features from surfaces ("texture"). Right: output of the texture system ("texture") to a real-world texture image ("luminance"). The last examples show that both lines and points are represented in the texture system.

Texture system. Texture is defined here as small-scale even symmetric contrast configurations. We distinguish two further subtypes: lines and points. Both subtypes are usually superimposed on surfaces (figure 4). Often, points are generated by noise. Therefore, by suitable interactions between the texture system and the surface system, noise may be discounted from object surface representations. The strength of this interaction may be modulated by an attentional system, since occasionally it may happen that points actually correspond to structure information. Nevertheless, no additional filtering (like a median filter) is required to achieve denoising, which is of particular interest for image processing. Above, we have briefly described the subsystems of our computational model. For image processing tasks, we now need to combine the output of all three subsystems. To do so, the output of the texture system is "printed" on the combined surface/gradient output. However, if we are interested in denoising tasks, we could eliminate the points, since the latter typically correspond to noise.

In order to combine the output of the surface system with the gradient system, preliminary simulations suggest that gradients should be built upon filled-in surface representations. These computational mechanisms, however, are subject of ongoing investigation.

3 Summary and Conclusions

Our model provides a novel view on early vision, since it emphasizes that the visual input should be interpreted by three subsystems accomplishing a segregation into surface, texture and gradient maps. In particular, with this segregation process we propose a new interpretation regarding the role of cortical simple cells in early vision, and their contribution to generate distributed representations of surface layout [15, 16]. We believe that this segregation facilitates object recognition, since it leads to separate intrinsic feature representations that are precursory to the generation of object surface representations. The gradient and texture system may provide additional information to higher visual areas. Unlike a simple coding approach that decomposes the visual input (e.g. [17]), we

propose a more richer representation that allows to semantically relate specific image content to underlying surface properties. Mechanisms which underly such a surface related processing necessitate more complex interactions in order to disambiguate and combine information from several maps.

References

1. Sepp, W., Neumann, H.: A multi-resolution filling-in model for brightness perception. In: ICANN99 Conference Publication. Volume 470., Ninth International Conference on Artificial Neural Networks (1999) 461–466
2. Arrington, K.: Directional filling-in. Neural Computation **8** (1996) 300–318
3. Pessoa, L., Mingolla, E., Neumann, H.: A contrast- and luminance-driven multiscale network model of brightness perception. Vision Research **35** (1995) 2201–2223
4. du Buf, J., Fischer, S.: Modeling brightness perception and syntactical image coding. Optical Engineering **34** (1995) 1900–1911
5. Neumann, H.: Mechanisms of neural architecture for visual contrast and brightness perception. Neural Networks **9** (1996) 921–936
6. McArthur, J., Moulden, B.: A two-dimensional model of brightness perception based on spatial filtering consistent with retinal processing. Vision Research **39** (1999) 1199–1219
7. Blakeslee, B., McCourt, M.: A multiscale spatial filtering account of the white effect, simultaneous brightness contrast and grating induction. Vision Research **39** (1999) 4361–4377
8. Grossberg, S., Todorović, D.: Neural dynamics of 1-d and 2-d brightness perception: A unified model of classical and recent phenomena. Perception & Psychophysics **43** (1988) 241–277
9. Gerrits, H., Vendrik, A.: Simultaneous contrast, filling-in process and information processing in man's visual system. Experimental Brain Research **11** (1970) 411–430
10. Knau, H., Spillman, L.: Brightness fading during ganzfeld adaptation. Journal of the Optical Society of America A **14** (1997) 1213–1222
11. Li, C.Y., Pei, X., Zhou, Y.X., von Mitzlaff, H.C.: Role of the extensive area outside the x-cell receptive field in brightness information transmission. Vision Research **31** (1991) 1529–1540
12. Li, C.Y., Zhou, Y.X., Pei, X., Qiu, F.T., Tang, C.Q., Xu, X.Z.: Extensive disinhibitory region beyond the classical receptive field of cat retinal ganglion cells. Vision Research **32** (1992) 219–228
13. Tittle, J., Todd, J.: Perception of three-dimensional structure. In Arbib, M., ed.: The Handbook of Brain Theory and Neural Networks. The MIT Press, Cambridge, Massachusetts (1995) 715–718
14. Ross, J., Morrone, M., Burr, D.: The conditions under which Mach bands are visible. Vision Research **29** (1989) 699–715
15. Hubel, D., Wiesel, T.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. Journal of Physiology-London **160** (1962) 106–154
16. Hubel, D., Wiesel, T.: Receptive fields and functional architecture of monkey striate cortex. Journal of Physiology-London **195** (1968) 214–243
17. Daugman, J.: Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. Journal of the Optical Socienty of America A **2** (1985) 1160–1169

Multi–Modal Statistics of Edges in Natural Image Sequences

Norbert Krüger and Florentin Wörgötter
University of Stirling, Scotland, norbert{worgott}@cn.stir.ac.uk

Abstract

In this work we investigate the multi-modal statistics of natural image sequences looking at the modalities orientation, color, optic flow and contrast transition. It turns out the statistical interdependencies corresponding to the Gestalt law collinearity increase significantly when we look not at orientation only

1 Introduction

A large amount of research has been focused on the usage of Gestalt laws in computer vision systems (overviews are given in [14, 13]). The most often applied and also the most dominant Gestalt principle in natural images is collinearity [3, 9]. Collinearity can be exploited to achieve more robust feature extraction in different domains, such as, edge detection (see, e.g., [7, 8]) or stereo estimation [2, 13]. In most applications in artificial visual systems, the relation between features, i.e., the applied Gestalt principle, has been defined heuristically based on semantic characteristics such as orientation or curvature. Mostly, explicit models of feature interaction have been applied, connected with the introduction of parameters to be estimated beforehand, a problem recognized as extremely awkward in computer vision. Recently, Geisler et al [6] introduced the idea to overcome heuristic and explicit models by relating feature interaction to the statistics of natural images. The feasibility of this approach becomes strong support from the *measurable interdependencies* of features in visual scenes that turn out to correspond to Gestalt laws [9, 3, 6].

In the human visual system beside local orientation also other modalities such as color and optic flow are computed (see, e.g. [5]). Gestalt principles are

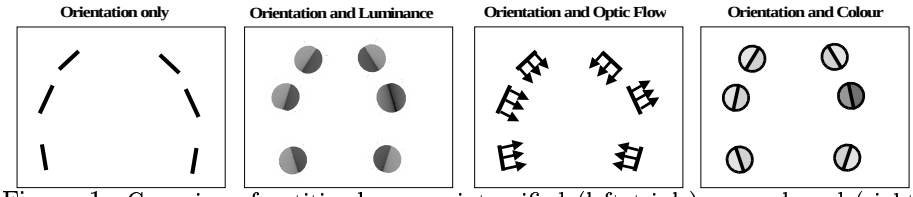


Figure 1: Grouping of entities becomes intensified (left triple) or weakened (right triple) by using additional modalities.

affected by multiple modalities. For example, figure 1 shows how collinearity can be intensified by the different modalities contrast transition, optic flow and color. This paper addresses statistics of natural images in these modalities. As a main result we found that statistical interdependencies corresponding to the Gestalt law "collinearity" in visual scenes become significantly stronger when multiple modalities are taken into account (see section 2).

2 Multi–Modal Statistics in Image Sequences

In the work presented here we address the multi–modal statistics of natural images. We start from a feature space (see also figure 1) containing the sub-modalities:
Orientation: We compute local orientation o (and local phase p) by the specific isotropic linear filter [4].
Contrast Transition: The contrast transition of the signal is coded in the phase p of the same filter.
Color: Color is processed by integrating over image patches in coincidence with their edge structure (i.e., integrating over the left and right side of the edge separately). Hence, we represent color by the two tuples $(c_r^l, c_g^l, c_b^l), (c_r^r, c_g^r, c_b^r)$ representing the color in RGB space on the left and right side of the edge.
Optic Flow: Local displacements (f_1, f_2) are computed by a well known optical flow technique ([11]).

2.1 Measuring Statistical Interdependencies:

We measure statistical interdependencies by the so called 'Gestalt coefficient' (see also [9]). The Gestalt coefficient is defined by the ratio of the likelihood of an event e^1 given another event e^2 and the likelihood of the event e^1:

$$G(e^1, e^2) = \frac{P(e^1|e^2)}{P(e^1).} \tag{1}$$

For the modeling of feature interaction a high Gestalt coefficient is helpful since it indicates the modification of likelihood of the event e^1 depending on other events. A Gestalt coefficient of one says, that the event e^2 does not influence the likelihood of the occurrence of the event e^1. A value smaller than one indicates a negative dependency: the occurrence of the event e^2 reduces the likelihood that e^1 occurs. A value larger than one indicates a positive dependency: the occurrence of the event e^2 increases the likelihood that e^1 occurs. The Gestalt coefficient is illustrated in figure 2. Further details can be found in [10].

2.2 Second Order Relations Statistics of Natural Images

A large amount of work has addressed the question of efficient coding of visual information and its relation to the statistics of images. Excellent overviews are given in [16, 15]. While many publications were concerned with the statistics on the pixel level and the derivation of filters from natural images by coding principles (see, e.g. [12, 1]), recently statistical investigation for local edge

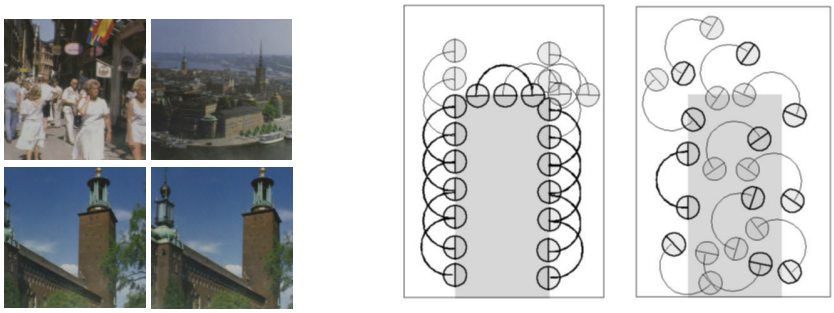


Figure 2: **Left:** Images of the data set (top) and 2 images of a sequence (bottom). **Right:** Explanation of the Gestalt coefficient $G(e^1|e^2)$: We define e^2 as the occurrence of a line segment with a certain orientation (anywhere in the image). Let the second order event e^1 be: "occurrence of collinear line segments two units away from an existing line segment e^2". Left diagram: Computation of $P(e^1|e^2)$. All possible occurrences of events e^1 in the image are shown. Bold arcs represent real occurrences of the specific second order relations e^1 whereas arcs in general represent possible occurrences of e^1. In this image we have 17 possible occurrences of collinear line segments two units away from an existing line segment e^2 and 11 real occurrences. Therefore we have $P(e^1|e^2) = 11/17 = 0.64$. Right diagram: Approximation of the probability $P(e^1)$ by a Monte Carlo method. Entities e^2 (bold) are placed randomly in the image and the presence of the event 'occurrence of collinear line segments two units apart of e^2' is evaluated. (In our simulations we used more than a 500000 samples for the estimation of $P(e^1)$). Only in 1 of 11 possible cases this event takes place (bold arc). Therefore we have $P(e^1) = 1/11 = 0.09$ and the Gestalt coefficient for the second order relation is $G(e^1|e^2) = 0.64/0.09 = 7.1$.

structures have been performed (see, e.g., [9, 3, 6]) and have addressed the representation of Gestalt principles.

Here we go one step further by investigating the second order relations not only in the modality orientation but in our multi–modal feature space

$$e = ((x_1, x_2), o, p, ((c_r^l, c_g^l, c_b^l), (c_r^r, c_g^r, c_b^r)), (f_1, f_2)).$$

In our simulations we collect second order events in bins defined by small patches in the (x_1, x_2)–space and by regions in the modality–spaces defined by the metrics defined for each modality (for details see [10]). Figure 3 shows the Gestalt coefficient for equidistantly separated bins (one bin corresponds to a square of 10×10 pixels and an angle of $\frac{\pi}{8}$ rad). As already been shown in [9, 6] collinearity can be detected as significant second order relation as a ridge in the surface plot for $\Delta o = 0$ in figure 3e. Also parallelism is detectable as an offset of this surface. A Gestalt coefficient significantly above one can also be detected for small orientation differences (figure 3d,f, i.e., $\Delta o = -\frac{\pi}{8}$ and $\Delta o = \frac{\pi}{8}$).

The general shape of surfaces is similar in all following measurements concerned with additional modalities: *we find a ridge corresponding to collinearity and an offset corresponding to parallelism and a Gestalt coefficient close to one for all larger orientation differences.* Therefore, in the following we will only look at the surface plots for equal orientation $\Delta o = 0$. These result shows *that Gestalt laws are reflected in the statistics of natural images: Collinearity and*

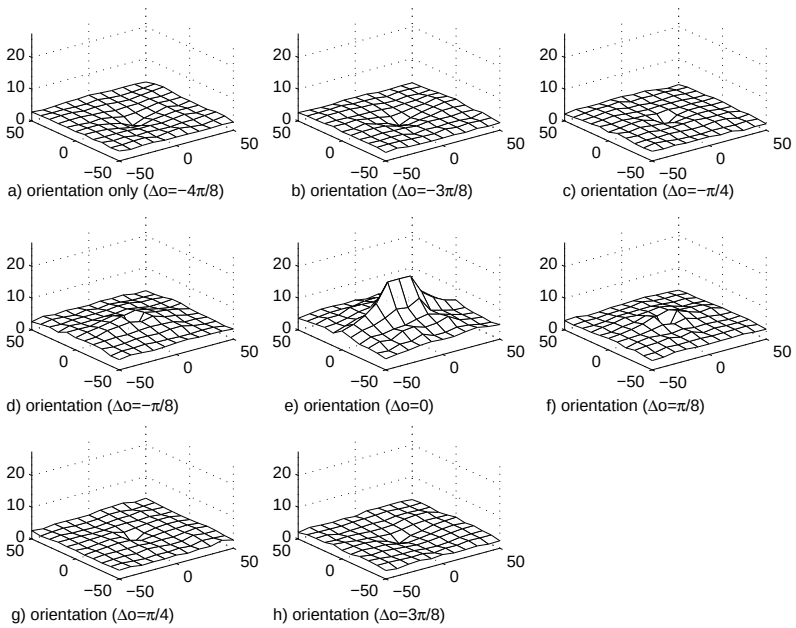


Figure 3: The Gestalt coefficient for differences in position from -50 to 50 pixel in x– and y– direction when orientation only is regarded. Note that the Gestalt coefficient for position (0,0) and $\Delta o = 0$ is set to the maximum of the surface for better display. The Gestalt coefficient is not interesting at this position, since e^1 and e^2 are identical

parallelism correspond to significant second order events of visual low level filters (see also [9]).

2.3 Pronounced Interdependencies by using additional Modalities

Now we can look at the Gestalt coefficient when we also take into account the modalities contrast transition, optic flow and color.

One additional modality: Figure 4b shows the Gestalt coefficient for the events 'similar orientation and similar contrast transition' (the metrics for the different modalities are defined precisely in [10]). In figure 5 the Gestalt coefficient along the x-axes in the surface plot of figure 4 is shown. The Gestalt coefficient on the x-axes correspond to the 'collinearity' ridge. The first column represents the Gestalt coefficient when we look at similar orientation only, while the second columns represent the Gestalt coefficient when we look at similar orientation and similar phase. *We see a significant increase of the Gestalt coefficient compared to the case when we look at orientation only corresponding to the Gestalt law collinearity.* Analogously, we define that two events have 'similar color structure' or 'similar optic flow'. The corresponding surface plot is shown in figure 4c and 4d. The slice corresponding to the collinearity ridge is shown in the third and fourth column in figure 5. An even more pronounced

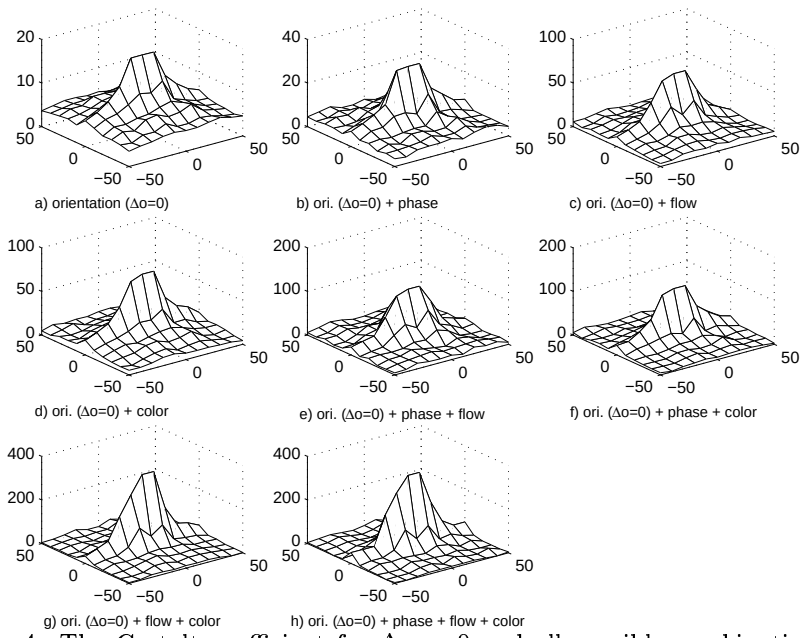


Figure 4: The Gestalt coefficient for $\Delta o = 0$ and all possible combination of modalities.

increase of inferential power for collinearity can be detected.

Multiple additional Modalities: Figure 4 shows the surface for similar orientation, phase and optic flow (figure 4e); similar orientation, phase and color (figure 4f) and similar orientation, optic flow and color (figure 4g). The slices corresponding to collinearity are shown in the fifth to seventh columns in figure 5. We can see that the the Gestalt coefficient for collinear line segments again increases significantly. Most distinctly for the combination optic flow and color (seventh column). Finally we can look at the Gestalt coefficient when we take all three modalities into account. Figure 4h and the eighth column in figure 5 shows the results. Again an increase of the Gestalt coefficient compared to the case when we look at only two additional modalities can be achieved.

Conclusion: In this paper we have addressed the statistics of local oriented line segments derived from natural scenes by adding information to the line segment concerning the modalities contrast transition, color, and optic flow. We could show that statistical interdependencies in the orientation–position domain correspond to the Gestalt laws collinearity and parallelism and that they become significantly stronger when multiple modalities are taken into account.

The results presented here provide further evidence for the assumption that despite the vagueness of low level processes stability can be achieved by *integration of information across modalities*. In addition, the attempt to model the application of Gestalt laws based on statistical measurements, as suggested recently by some researchers (see, [6, 3, 9]) gets further support. Most importantly, the results derived in this paper suggest to formulate the application of Gestalt principles in a multi-modal way.

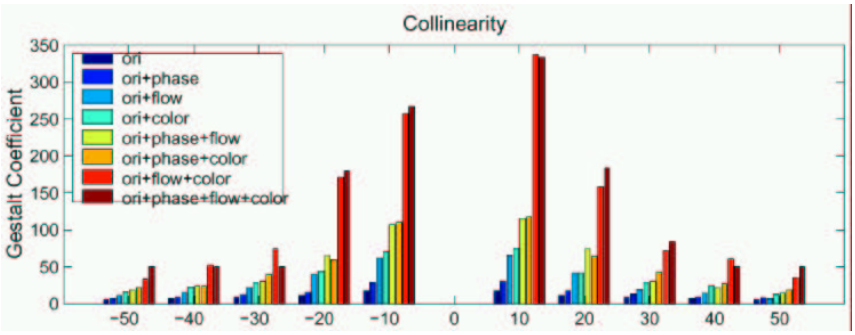


Figure 5: The Gestalt coefficient for collinear feature vectors for all combinations of modalities. The x-axis represents the distance of the collinear line segments in pixel and corresponds to the collinearity ridge in figure 3 and 4. For (0,0) the Gestalt coefficient is not shown, since e^1 and e^2 would be identical.

References

[1] A.J. Bell and T. Sejnowski. Edges are the 'independent components'of natural scenes. *Advances in Neural Information Processing Systems*, 9, 1996.

[2] R.C.K. Chung and R. Nevatia. Use of monucular groupings and occlusion analysis in a hierarchical stereo system. *CVPR*, 1991.

[3] H. Elder and R.M. Goldberg. Inferential reliability of contour grouping cues in natural images. *Perception Supplement*, 27, 1998.

[4] M. Felsberg and G. Sommer. The monogenic signal. *IEEE Transactions on Signal Processing*, 41(12), 2001.

[5] M.S. Gazzaniga. *The cognitive Neuroscience*. MIT Press, 1995.

[6] W.S. Geisler, J.S. Perry, B.J. Super, and D.P. Gallogly. Edge co–occurrence in natural images predicts contour grouping performance. *Vision Research*, 41:711–724, 2001.

[7] G. Guy and G. Medioni. Inferring global perceptual contours from local features. *International Journal of Computer Vision*, 20:113–133, 1996.

[8] F. Heitger, R. von der Heydt, E. Peterhans, L. Rosenthaler, and O. Kübler. Simulation of neural contour meachnisms: representing anomalous contours. *Image and Vision Computing*, 16:407–421, 1998.

[9] N. Krüger. Collinearity and parallelism are statistically significant second order relations of complex cell responses. *Neural Processing Letters*, 8(2):117–129, 1998.

[10] N. Krüger and F. Wörgötter. Multi modal estimation of collinearity and parallelism in natural image sequences. *to appear in Network: Computation in Neural Systems*, 2002.

[11] H.-H. Nagel. On the estimation of optic flow: Relations between different approaches and some new results. *Artificial Intelligence*, 33:299–324, 1987.

[12] B.A. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[13] S. Posch. *Perzeptives Gruppieren und Bildanalyse*. Habilitationsschrift, Universität Bielefeld, Deutscher Universitäts Verlag, 1997.

[14] S. Sarkar and K.L. Boyer. *Computing Perceptual Organization in Computer Vision*. World Scientific, 1994.

[15] E.P. Simoncelli and B.A. Ohlshausen. Natural image statistics and neural representations. *Anual Reviews of Neuroscience*, 24:1193–1216, 2001.

[16] C. Zetsche and G. Krieger. Nonlinear mechanisms and higher–order statistics in biologial vision and electronic image processing: review and perspectives. *Journal of electronic imaging*, 10(1), 2001.

Inferring Salient Features in Images by Perceptual Grouping with Inhibitory and Excitatory Tensor Fields

Amin Massad and Bärbel Mertsching *

University of Hamburg, Dept. of Computer Science, IMA-Lab
Vogt-Kölln-Str. 30, D-22527 Hamburg, Germany
massad@informatik.uni-hamburg.de

Abstract. We present the extension of the perceptual grouping technique known as *Tensor voting* to the application to grey-level images. The image data is encoded by a tensorial representation of the local orientation which is computed from a set of Gabor filters. The resulting dense tensor maps are refined by means of newly introduced inhibitory voting fields. Subsequent grouping with excitatory voting fields yields saliency maps for contours and junctions.

1 Overview

We present a perceptual grouping approach applicable to images with the aim to facilitate a transition from local low-level features to more global high-level information. The method allows the inference of salient contours and junctions from images based on the principles of good continuation and proximity, which according to psychological studies [4, 5] play a special role among the set of Gestalt laws.

By the computation of local orientation tensors from a set of Gabor filters, our approach extends the *tensor voting* (TV) technique developed by [9] to the application to grey-level images. Using second order tensors as input and output tokens, we simultaneously encode information about orientation and orientation uncertainty – in contrast to other vector-based grouping methods which can only represent direction (e. g. [1, 3, 10–14]). Other advantages of the method are the exclusive use of local operations and its linearity. Moreover, due to the similarity to a convolution operation, computation does not involve iterative processing as required in other optimization-like approaches.

While inputs formerly consisted of binary images or sparse edgel maps, our extension yields oriented input tokens and the locations of junctions as input to the perceptual grouping. In order to handle dense input maps, the tensor voting framework is extended by the introduction of grouping fields with inhibitory regions.

* We gratefully acknowledge partial funding of this work by the Deutsche Forschungsgemeinschaft under grant Me1289/7-1 “KomForm”.

2 Tensor Voting

For a brief review of the TV framework, we restrict our explanations to the 2D-case where a second order symmetric tensor over $\mathbb{R}^2$ can be denoted by a symmetric 2×2 matrix $T = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^\top + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2^\top$ with two perpendicular eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2$ and two corresponding real eigenvalues $\lambda_1 > \lambda_2$. Basically, the tensor represents the second order moments of the local orientation for each image location and can be visualized by an ellipse.

The definition of saliency measures is deducted from the decomposition of a tensor into

$$T = (\lambda_1 - \lambda_2)\boldsymbol{e}_1 \boldsymbol{e}_1^\top + \lambda_2(\boldsymbol{e}_1 \boldsymbol{e}_1^\top + \boldsymbol{e}_2 \boldsymbol{e}_2^\top)\,. \tag{1}$$

In (1), the weighting factor $(\lambda_1 - \lambda_2)$ represents an orientation certainty in the direction of the eigenvector $\boldsymbol{e}_1$ and thus will be called *curve-* or *stick-saliency*. The second weight λ_2 is applied to a circle, thus we call it *junction-* or *ball-saliency* because it indicates a high orientation uncertainty which is equivalent to the confidence in the presence of a junction.

Figure 1a illustrates that the tensor addition of similarly oriented tensors yields an increased stick-saliency whereas differently oriented tensors yield a high ball-saliency.

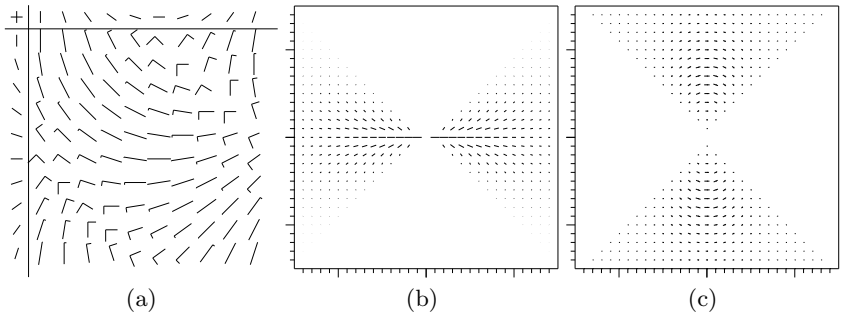


Fig. 1. (a) Tensor addition: The tensors are depicted by $\lambda_1 \boldsymbol{e_1} \perp \lambda_2 \boldsymbol{e_2}$. (b) Excitatory stick-voting field for a horizontally oriented input token P at the center. (c) Inhibitory stick-voting field.

Grouping is achieved by the interaction of input tokens according to their stick-saliency or ball-saliency, respectively. In the case of oriented input tokens, stick-voting is applied: For each token the stick-voting-field (Fig. 1b) is aligned to its eigenvector $\boldsymbol{e}_1$ and weighted with $\lambda_1 - \lambda_2$ and all fields are combined in a convolution-like manner by tensor addition. The layout of this field encodes the connection of neighboring tokens which fulfill the minimal curvature constraint. Hence, it allows to strengthen locally collinear or co-circular structures, including virtual contours across gaps in the image data.

3 From Local Orientation to Tensor Tokens

In order to apply the TV technique to grey-level images, we transform the image data into a tensor description. This is achieved by the computation of the local orientation and orientation certainty from a set of quadrature filters. We use two-dimensional Gabor filters for their known optimality with regards to the time-bandwidth product:

$$g(\boldsymbol{k}) = K \exp\left(-\frac{1}{2}(\boldsymbol{k}-\boldsymbol{k}_0)^\top D(\boldsymbol{k}-\boldsymbol{k}_0)\right) \tag{2}$$

where $\boldsymbol{k}$ denotes the frequency, K a normalization constant and D a 2×2 covariance matrix. The kernel consists of a two-dimensional Gaussian centered around $\boldsymbol{k}_0$ with variances σ_1^2, σ_2^2 as the eigenvalues of D. A similar approach, but with different filter kernels, has been used by [2].

The response $g_i(\boldsymbol{x})$ of Gabor filter i, with the center frequency $\boldsymbol{k}_{0,i}$ at image position $\boldsymbol{x}$, is a measure for orientation certainty in the direction of that filter. Therefore, we introduce the orientation tensor $T_i = \boldsymbol{e}_i \boldsymbol{e}_i^\top$, which represents an ideal orientation in the direction of the unit vector $\boldsymbol{e}_i$ perpendicular to $\boldsymbol{k}_{0,i}$. Then, the weighted tensor sum

$$T(\boldsymbol{x}) = \sum_{i=1}^{n} g_i(\boldsymbol{x}) T_i \tag{3}$$

over all filter orientations i gives an estimate for the local orientation and orientation uncertainty at image position $\boldsymbol{x}$. Figure 2 shows the tensors which result from applying this procedure to the image of a circle. Note that the locations along the contour with higher orientation uncertainty correspond to alias effects. They are caused by the discretization of the image and detected in dependence of the parametrization of the Gabor filters.

In order to facilitate the inference of image features larger than the Gabor kernel size, the voting field size σ_v is a function of the Gabor kernel size σ_g: The relation $\sigma_v/\sigma_g = 6$ is derived from results of psychophysical experiments [1].

4 Inhibitory Voting Fields

Initially, TV has been designed to group sparse input maps by means of a densification in order to identify m-D structures in n-D input space with $m < n$ (i. e. lines and points in 2-D space). However, due to the localization uncertainty of Gabor filters, the Gabor transform yields for 0-D or 1-D image structures input tensors which extend over regions. In order to compensate for this blurring effect and to fit the input tokens better to the model of the voting field design, we have proposed to apply a non-maximum suppression method to the local orientation tensors prior to the grouping process. This thinning step has been embedded into the TV framework by the introduction of inhibitory voting fields, please refer to [7] for additional details of the algorithm.

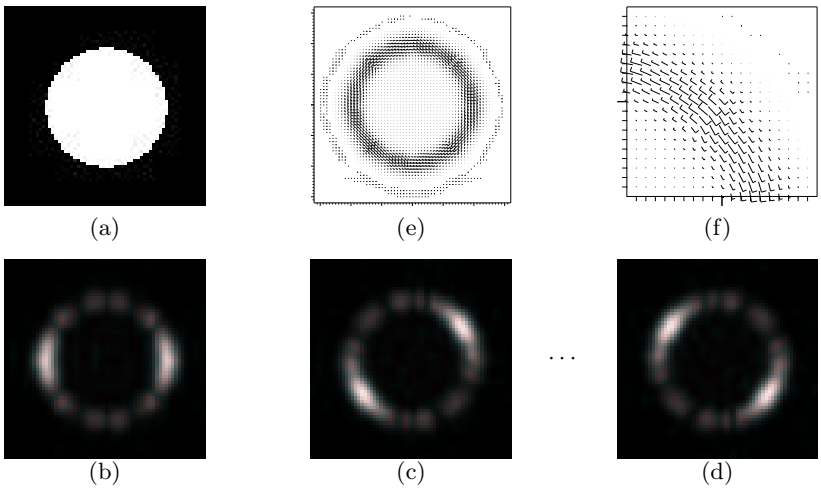


Fig. 2. From local orientation to tensors. (a) Input image. (b)-(d) Filter responses $g_0(\boldsymbol{x}), g_1(\boldsymbol{x}), \ldots, g_4(\boldsymbol{x})$. (e) Tensor map $T(\boldsymbol{x})$. (f) Upper right quarter of (e) zoomed by a factor of 4.

The inhibitory voting field is designed to operate on areas complementary to the excitatory voting field: The excitatory stick-voting field proposed by [9] (Figure 1b) only covers the region F_+ and leaves out the region F_- with $\frac{\pi}{4} \leq \theta \leq \frac{3}{4}\pi$. The region F_- is excluded from excitatory grouping because the assumed circular connection with an *oriented* input token at P does not fulfill the minimal total curvature constraint (an elliptic connection would yield lower total curvature).

The inhibitory voting field (Fig. 1c) covers exactly these complementary positions F_-, which have previously been excluded from the grouping process. This newly defined field achieves edge thinning by suppressing orientations which are approximately parallel to an oriented input token P and have lower saliencies $sal(Q) < sal(P)$. Because non-maxima locations are assumed to lie perpendicular to the orientation of P, inhibition should be strongest at angles $\theta \approx \frac{\pi}{2}$ where $Q \parallel P$ and decrease to zero towards the two extremal cases along the circle $\theta \approx \frac{\pi}{4}$ and $\theta \approx \frac{3}{4}\pi$ where $Q \perp P$.

The strength of the inhibition is defined as

$$F_-(r,\theta) = \begin{cases} sal(P) \cdot \left(e^{-\frac{1}{2}\frac{r^2}{\sigma_1^2}} - e^{-\frac{1}{2}\frac{r^2}{\sigma_2^2}} \right) \cdot \cos^8(\theta) & \text{if} \quad \frac{\pi}{4} \leq |\theta| \leq \frac{3}{4}\pi \\ 0 & \text{else} \end{cases} \tag{4}$$

which is an adaptation of the formula by [3] overlaid with a difference of Gaussians (with $\sigma_1 > \sigma_2$ to model an off-surround behavior, while the on-center part consists of the excitatory field). The orientations $\boldsymbol{e}(r,\theta)$ of the field tokens are defined by the normalized tangent vectors of the circles cotangent to P and encoded as stick-tensors $T = F_-(r,\theta) \cdot \boldsymbol{e}\boldsymbol{e}^\top$.

5 Results

Figure 3 gives an example of an image where the application of Gabor filters is not sufficient to extract salient structures. In order to bridge gaps and to compensate for considerably high noise, grouping is needed to infer structures beyond the size of a Gabor kernel.

Salient contours (Figure 3e) are extracted from saliency maps by the application of an adapted marching squares algorithm which traces the contours along maximal saliencies and yields a subpixel-accurate vectorial representation of the curve. By means of this method, it becomes possible to compute the positional precision of contours and junctions which is subject to ongoing research.

In contrast to [6], our approach infers salient structures based on local operations compared to global connections between all image features. Grouping is based on the principles of good continuation and proximity and does not require further assumptions about the objects' geometry. Moreover, the computation of local orientation tensors doesn't hypothesize step-edges, which isn't valid at corners, but rather represents them as locations with high orientation uncertainty.

References

1. D. Field, A. Hayes, and R. Hess. Contour integration by the human visual system: Evidence for a local association field. *Vision Research*, 33:173–193, 1993.
2. Gösta Granlund and Hans Knutsson. *Signal Processing for Computer Vision.* Kluwer Academic Press, 1995.
3. F. Heitger and R. v. d. Heydt. A computational model of neural contour processing: Figure-ground segregation and illusory contours. In *ICCV*, pages 32–40, 1993.
4. P. Kellman and T. Shipley. A theory of visual interpolation in object perception. *Cognitive Psychology*, 23:141–221, 1991.
5. I. Kovacs. Gestalten of today: Early processing of visual contours and surfaces. *Behav. Brain Res.*, 82:1–11, 1996.
6. S. Mahamud, K. Thornber, and L. Williams. Segmentation of salient closed contours from real images. In *ICCV*, volume 2, pages 891–897, 1999.
7. A. Massad, M. Babos, and B. Mertsching. Application of the tensor voting technique for perceptual grouping to grey-level images. In *Pattern Recognition 24th DAGM Symposium*, 2002.
8. A. Massad, M. Babos, and B. Mertsching. Perceptual grouping in grey level images by combination of gabor filtering and tensor voting. In *ICPR*, volume 2, pages 677–680, 2002.
9. G. Medioni, M. Lee, and C. Tang. *A Computational Framework for Segmentation and Grouping.* Elsevier, 2000.
10. M. Nitzberg and D. Mumford. The 2.1-D sketch. In *ICCV*, pages 138–144, 1990.
11. S. Sarkar and K. Boyer. Integration, inference, and managment of spatial information using bayesian networks: Perceptual organization. *PAMI*, 15:256–274, 1993.
12. E. Saund. Perceptual organization of occluding contours of opaque surfaces. *CVIU*, 76:70–82, 1999.
13. A. Sha'ashua and S. Ullman. Structural saliency: the detection of globally salient structures using a locally connected network. In *ICCV*, pages 312–327, 1998.
14. K. Thornber and L. Williams. Analytic solution of stochastic completion fields. *Biol. Cybern.*, 75:141–151, 1996.

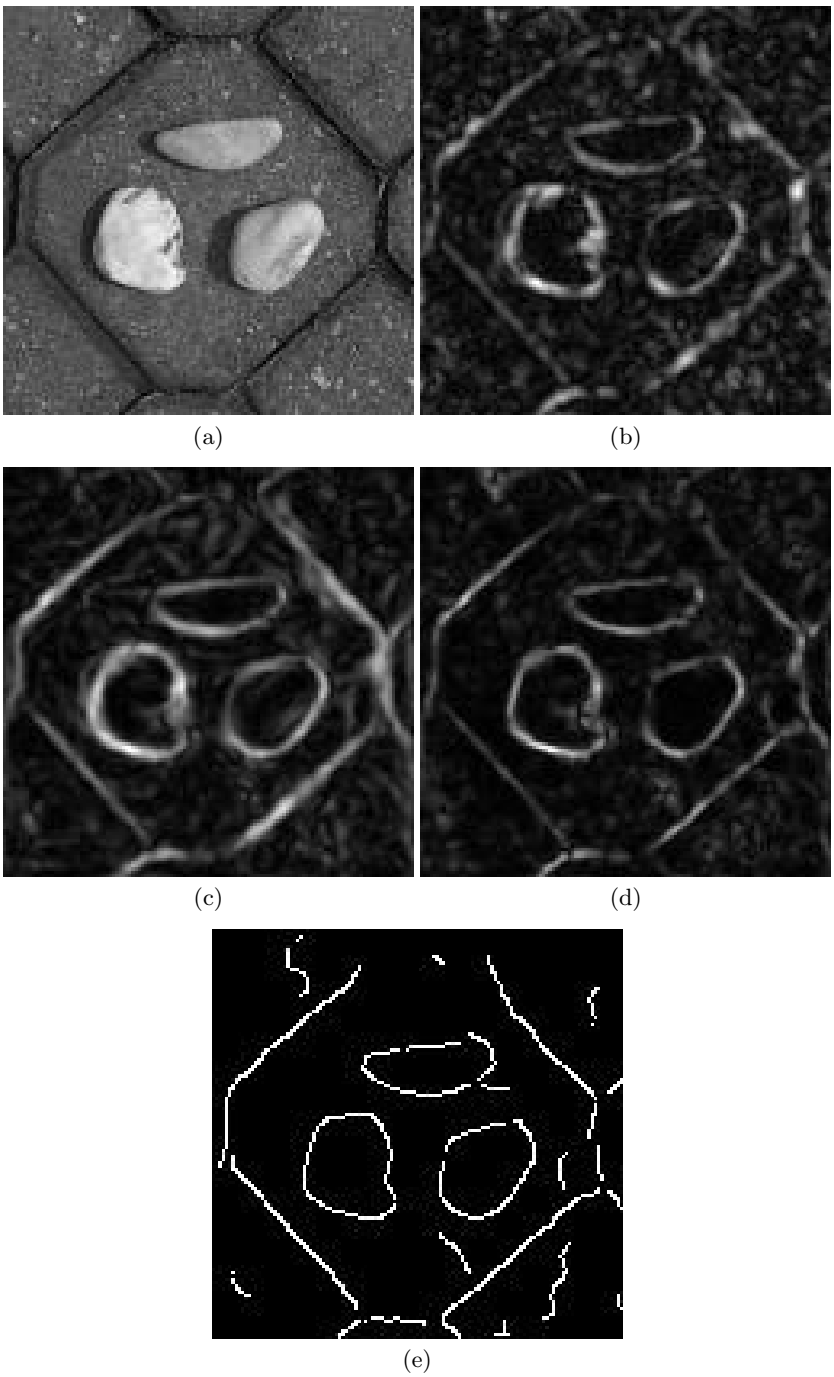


Fig. 3. Results on an natural scene: (a) Input image from [6]. (b) Stick-saliency of Gabor responses. (c) Stick-saliency with excitatory voting only, as in [8]. (d) Stick-saliency from combination of inhibitory and excitatory voting. (e) Salient curves extracted from (d) by application of a marching squares algorithm.

The Statistics of Natural Scenes and Weber's Law

Florian Röhrbein and Christoph Zetzsche

Institut für Medizinische Psychologie
Ludwig-Maximilians-Universität München, Goethestr. 31, 80336 München
{florian,chris}@imp.med.uni-muenchen.de

Abstract. The classical linear filtering properties of early vision can be explained as an information-theoretically optimized adaptation to the statistical dependencies in natural scenes. Here we investigate whether a fundamental nonlinear property of visual perception, Weber's law, can also be explained in such a statistical framework. We measure the joint statistics of neighbouring pixels of natural images under varying illumination conditions, and demonstrate that a linear decorrelating transform by DOG filters would leave significant statistical dependencies between the responses. We then show that the removal of these statistical dependencies requires a nonlinear gain control mechanism which can be implemented as ROG (ratio of Gaussian) filter. Weber's law is a direct consequence of this nonlinear operation. A single principle, the reduction of statistical dependencies between sensory messages, is thus sufficient to derive all essential processing properties of early vision.

1 Introduction

Recent investigations of the statistics of natural scenes and of their neural representations have indicated that the decomposition by size- and orientation-selective filters can be explained as an information-theoretically optimized adaptation to the statistical redundancies of the natural environment (for review see, e.g., [1]). More recent developments indicate that this approach can also be extended to more complicated cortical processing properties, as in complex cells or in the extra-classical receptive field surround [2][3][4]. According to the information-theoretic approach the neural operations represent a transformation of the state space coordinates which matches the representation to the structure of the multivariate probability distribution. One major criterion for a good match (though not the only one) is the reduction of the statistical dependencies (ideally: *statistical independence*). Often this can be achieved by linear transforms, as in independent component analysis (ICA), but some statistical dependencies require nonlinear operations. Here we investigate whether the nonlinear operation underlying Weber's law can be seen as an efficient first step towards the goal of statistical independence.

2 Natural Scene Statistics and Linear Filter Decompositions

For this, we first measured the joint statistical distribution of the responses of neighbouring retinal receptors to natural scenes with varying lightning conditions (Fig. 1).

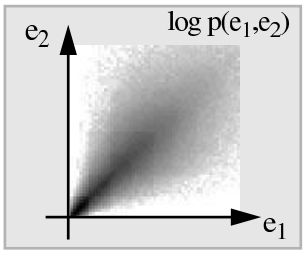


Fig. 1. Joint two-dimensional probability density function (pdf) of neighboring pixels in scenes with spatially and/or temporally varying illumination. 12 images were randomly taken from the van Hateren database of natural images [5] under exclusion of non-natural objects or portions of sky. Before the statistics were computed, the images were converted to an absolute intensity scale by a mapping which takes into account the aperture and the exposure time used in recording each image. The resulting pdf exhibits a typical shape: a high correlation between the pixel values and a systematic outward widening of the distribution towards the higher intensity values

A linear decorrelating transformation, like a multi-scale bandpass filtering by a difference of gaussian pyramid (DOG pyramid) [6], can exploit the statistical second-order dependencies of these joint statistics. To obtain a simple measure of the joint statistics of the DOG responses we can decompose the DOG pyramid for resolution level i as

$$g_i = \underbrace{(g_i - g_{i+1})}_{d_i} + \underbrace{(g_{i+1} - g_{i+2}) + (g_{i+2} - g_{i+3}) + \ldots + (g_{n-1} - g_n) + (g_n)}_{\Sigma d_{i+1}}. \quad (1)$$

Here d_i denotes a DOG channel with resolution i, and Σd_{i+1}(in slight abuse of notation) denotes the sum of the lower frequency DOG channels, which is equivalent to the local mean (i.e., $\Sigma d_{i+1}=g_{i+1}$). The decorrelation by a DOG representation can then be seen as "rotation" of the coordinate system (Fig. 2).

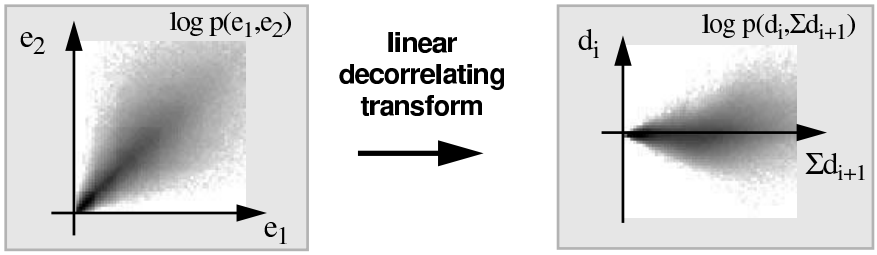


Fig. 2. Decorrelation of the joint statistics by a set of linear DOG-filters

However, this linear transformation cannot provide a separation of the higher-order dependencies which are reflected in the systematical dependence of the variance σ^2 of the DOG-response d_i on Σd_{i+1} (i.e., on the set of DOG channels with lower frequencies, which is represented by the local mean g_{i+1}). The true structure of the multivariate pdf of natural images is thus not separable in *linear* Cartesian coordinates (e.g., by PCA or ICA), but requires a nonlinear transformation.

This deficit of linear decorrelation is not surprising given the fact that the retinal input E(x) results from a nonlinear, multiplicative combination of an illumination component I(x) and an reflectance component R(x), i.e. E(x)=I(x)R(x). If various different illumination functions $I_k(x)$ occur across space and time, this will cause that one and the same reflectance function R(x) is transformed into different luminance functions $E_k(x)$. The corresponding statistical contributions $p_k(E_k(x))$ are scaled versions of each other, and constitute together the pdf p(E(x)). This effect is illustrated in Fig. 3.

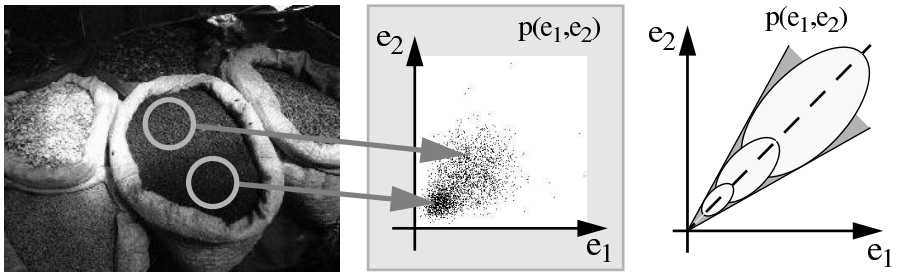


Fig. 3. The left image shows a typical configuration with spatially varying illumination. The grain in the sack is illuminated partially by direct bright light, and partially by indirect dim light. This gives rise to two contributions to the pdf, which are scaled versions of one another. Their combination constitutes the total joint pdf of neighbouring pixels in the grain region (center). The principle is schematically illustrated in the right figure

The combination of various scaled subpopulations in the final pdf is the reason for the statistical dependency of the DOG response d_i on the mean that has been revealed in Fig. 2. Let us hence take a closer look at this insufficiency of linear decorrelation schemes. The crucial point is that a linear decomposition cannot separate the nonlinear interaction of the reflectance component and the illumination component (Fig. 4).

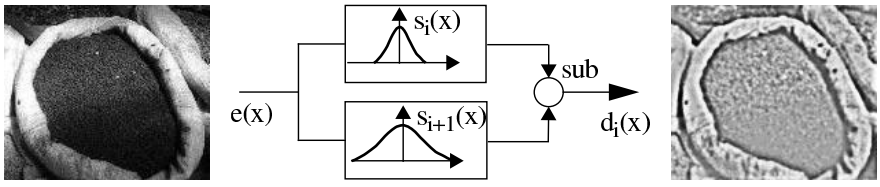


Fig. 4. Illumination dependence of a linear DOG operator. The response (right image) is proportional to the linear local differences in the image. The response to the grain texture is hence smaller in the dimly lit region than in the bright region, i.e. it is dependent on the illumination

3 Nonlinear Removal of Statistical Dependencies

Illumination changing often gradually across space, it is commonly assumed that a mere suppression of the low frequencies by a linear band-pass filter is already sufficient for the provision of illumination invariance. This is not the case, since a linear filter response will still be contaminated by the influence of the illumination. For one and the same reflectance structure, it will take greater values in the directly illuminated areas then in the dimly lit areas. This raises the question whether it is possible to find a suitable nonlinear transformation to get rid of these dependencies.The crucial factor that causes the statistical dependencies is the *proportionality* of the filter response *to the local mean*. A suitable nonlinear transform has thus to get rid of this proportionality, i.e. it has to *reduce* the gain of the system *in proportion to the local mean*. This can be achieved by an adaptive gain control mechanism. A straightforward realization of such a mechanism is a divisive interaction by a "ratio of Gaussians" (ROG) operator [7][8][9][10]. (A logarithmic transducer function would have a similar effect but would be much less suited for the processing of a wide dynamic range). The effect of this adaptive nonlinear operation is illustrated in Fig. 5.

Due to this nonlinear separation capabilities, such a ROG operator can in fact avoid the statistical dependency of the operator response on the local mean that has been observed for the linear DOG operator. The neural representation thus comes substantially closer to the desired statistical independence (Fig. 6).

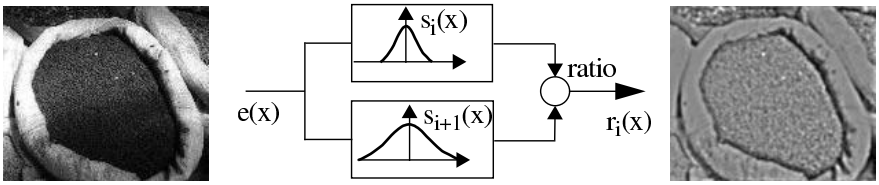


Fig. 5. Processing of illumination effects by a nonlinear ROG operator. The operator can separate the influence of the illumination by responding only to the reflectance (which is a homogeneous texture of grain in this example). As a consequence, the response will be statistically independent of the illumination (cf. Fig. 4)

Fig. 6. The joint statistics of the response r_i of a ROG operator and of the set of remaining channels Πr_{i+1} are much closer to statistical independence (cf. Fig. 2) (note that in analogy to eq. (1) $\Pi r_{i+1}=\Sigma d_{i+1}=g_{i+1}$, i.e it is also equivalent to the local mean)

The nonlinear transformation that is required for the removal of the statistical dependencies yields Weber's law as a direct consequence. It produces an input-output relation in which a fixed output increment is caused by all those inputs for which the input increment ΔI is proportional to the local mean I_0. This is just Weber' s law $\Delta I/I_0$=const. (Fig. 7).

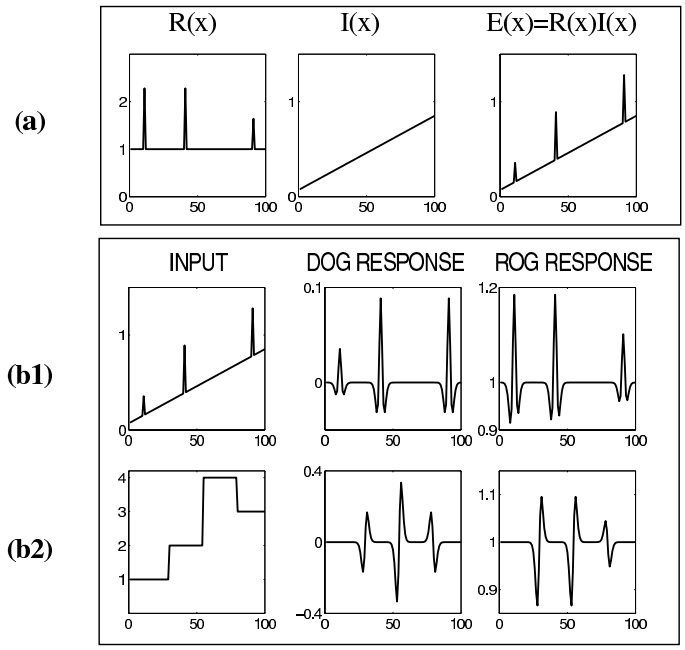


Fig. 7. ROG response and Weber's law. (a) The input E(x) results from a nonlinear, multiplicative combination of the reflectance R(x) and the illumination I(x). (b1) A linear DOG operator cannot separate these components since it responds in proportion to the linear signal differences. The ROG transform causes constant response increments for input ratios $\Delta I/I_0$=const., i.e., yields Webers law. By this, it can separate the reflectance component from the illumination component. (b2) A second example with step inputs

4 Conclusion

Weber's law describes a fundamental nonlinearity of the visual system. It is often used in a purely descriptive manner, without explicit reference to an underlying functional basis. If a function is considered, then it is associated to the perceptual invariance of lightness constancy. For this it is assumed that during the course of evolution the visual system has somehow acquired a sort of knowledge about the physical laws of image formation, and that this knowledge has been incorporated into the neural processing to obtain an illumination-invariant characterization of physical objects. The present investigation suggests a much simpler way of development: Substantial statistical

redundancies are a typical characteristic of sensory messages, and the exploitation of such redundancies seems to be a universal strategy for an efficient representation of sensory information. Our statistical analysis has shown that there exist significant statistical dependencies between early sensory signals which, contrary to common assumptions, cannot be exploited by classical linear decorrelation schemes. The nonlinear transformation that is required to eliminate these dependencies yields illumination invariance and Weber's law. Like other basic visual functions, Weber's law can thus be seen as a consequence of one single principle: the visual system seeks to exploit the statistical redundancies of natural scenes.

Acknowledgment: Supported by DFG (SFB 462, GRK 267). We thank U. Nuding for programming assistance, and G. Hauske and I. Rentschler for helpful comments.

References

[1] C. Zetzsche and G. Krieger. Nonlinear mechanisms and higher-order statistics in biological vision and electronic image processing: review and perspectives. *J. Electronic Imaging*, 10(1):56–99, 2001.

[2] A. Hyvarinen and P. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

[3] O. Schwartz and E. Simoncelli. Natural signal statistics and sensory gain control. *Nat. Neurosci.*, 4(8):819–825, 2001.

[4] C. Zetzsche and F. Röhrbein. Nonlinear and extra-classical receptive field properties and the statistics of natural scenes. *Network: Comput. Neural Syst.*, 12:331–350, 2001.

[5] J. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. B*, B 265: 359-366, 1998.

[6] P. J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. *IEEE COM*, 31(4):532–540, April 1983.

[7] H. Wallach. Brightness constancy and the nature of achromatic colors. *J. Exp. Psychol.*, 38:310–324, 1948.

[8] G. Sperling. Model of visual adaptation and contrast detection. *Perception and Psychophysics*, 8:143–157, 1970.

[9] C. Zetzsche and G. Hauske. Multiple channel model for the prediction of subjective image quality. *Proc. SPIE*, 1077: 209–216, SPIE, Bellingham, WA, 1989

[10] L. Pessoa, E. Mingolla, and H. Neumann. A contrast- and luminance-driven multiscale network model of brightness perception. *Vision Res.*, 35(15):2201–2223, 1995.

The Visual System is "Blind" to Almost All Possible Images

Christoph Zetzsche

Institut für Medizinische Psychologie
Ludwig-Maximilians-Universität München, Goethestr. 31, 80336 München
chris@imp.med.uni-muenchen.de

Abstract. The operations in early vision can be seen as result of an optimal adaptation to the statistical redundancies of natural scenes. This hypothesis is mainly supported by the analysis and simulation of the statistical properties of single neurons, whereas the evidence is less clear on the behavioral level. Here we show that basic visual functions, like the discrimination of two images, do only work properly for natural images, whereas they suffer a complete breakdown for non-natural images (i.e., for images that lack any of the characteristic statistical redundancies of the natural ones, like random images or images with artificial redundancies). Since it can be formally proven that almost all possible images belong to this latter group, this implies that the visual system is specialized for the processing of a tiny fraction of the possible images, whereas it is functionally "blind" to almost all possible images.

1 Introduction

The hypothesis of an optimal adaptation of the early visual processing stages to the statistical redundancies of natural scenes has recently received increasing attention (for review see, e.g., [1]). Most arguments in support of this hypothesis have been derived from an analysis of the processing properties of individual neurons in the retina or in the visual cortex. Regarding behavioral properties, the situation is more complicated. Here the straight-forward prediction from the adaptation hypothesis would be that visual perception should work best within the class of natural images, whereas performance should be substantially reduced for the class of non-natural images (Fig. 1).

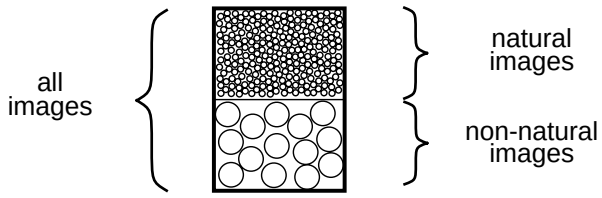


Fig. 1. Schematic illustration of the predicted visual discrimination capabilities (the "granularity" of the just noticeable differences). Fine discriminations should only be possible for natural images, whereas non-natural images should only allow for coarse discriminations

Recent investigations of this prediction employed manipulations of the statistical second-order properties. The analysis became complicated because for the class of natural images there is both evidence for a maximum sensitivity, but also for a greater robustness against changes (i.e. for a reduced sensitivity), depending on the experimental technique [2][3]. This prompted us to search for a simple and straightforward behavioral test of the adaptation hypothesis. Surprisingly, there exists a very simple test, that can provide considerable insight, but has not yet received the attention it deserves.

2 Perception of Random Images

In comparing the perceptual performance for natural vs. non-natural images, it seems not immediately obvious how the non-natural images should be designed. However, a prototypical design results from the basic property that characterizes natural images in terms of information theory: their high degree of structural regularity, or statistical redundancy. The prototypical non-natural images are thus *random images*, since these lack any of the statistical regularities of the natural ones. The behavioral test then becomes a simple issue (Fig. 2).

The observable complete break-down of visual discrimination (and thereby of any higher visual function, such as pattern recognition or classification) within the class of random images is a massive and well known effect, but for some reason it is usually not considered very relevant. This is a mistake, however, as will be shown in the following. The underestimation of the theoretical importance of the effect is presumably a direct consequence of our perceptual properties. We have the strong subjective impression of a high similarity and homogeneity of the random images, as opposed to the complexity and wide diversity of natural images, and we tend hence to believe that there must exist many more different natural images than different random images. However, in both cases we are fooled by our perceptual system. The physical difference of two random images is typically just as large as the difference between two arbitrary natural images. In fact, the state space vectors of two sample random images are in almost all cases close to orthogonal (Fig. 3). Even more dramatic is our perceptual misguidance with respect to the large number of images that we expect to find in the natural set, as opposed to those in the random set.

To understand this, we have to take a look at the theoretical approaches to randomness. From a probabilistic perspective, a sequence can be declared random if it is a

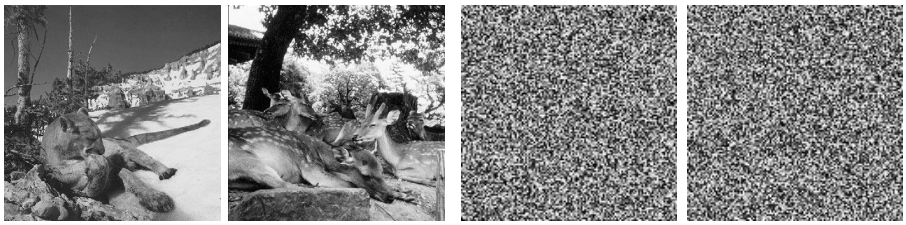


Fig. 2. Natural images can be easily distinguished, random images not

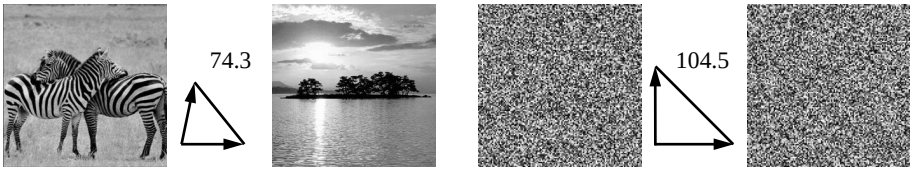


Fig. 3. Random images are as different from one another as are natural images. The dissimilarity of the images, as measured in terms of Euclidean distance, is indicated by the vectors diagrams

"typical" outcome and passes all conceivable tests of randomness. For example, it has to be Borel normal, i.e. all letters and blocks of letters should appear with approximately equal frequency. This can be formalized as universal Martin-Löf test [4].

How many of all the possible different sequences are then random? Surprisingly many, because, simply speaking, the number of typical sequences, i.e. of possible combinations with approximately equidistributed letters, increases rapidly with increasing size of the sequences (Fig. 4). Formally, the set of random infinite sequences has uniform measure one [4].

That almost all possible sequences are random can also be deduced from the different perspective of algorithmic complexity [5][6], which defines randomness as incompressibility: the shortest program that can describe a random string is not allowed to be significantly shorter than the string itself. Consider a simple counting analysis for binary strings: There are 2^n different strings of length n. A string is declared random if its shortest description has length greater m, with m only by a negligible fraction smaller than n. There exist 2^m different descriptions of length m, 2^{m-1} of length m-1, ..., 2 descriptions of length 1. Hence there exist in total 2^{m+1}-2 different descriptions with length no longer than m, and therefore at most a fraction of $(2^{m+1}-2)/2^n \approx 1/2^{n-m-1}$ of all strings of size n can be not random. Let us declare a binary image of size n=512x512 as random if its shortest description is longer than $0.9999 \cdot n$. This implies that 99.99999% of all images of this size are random.

In conclusion, whatever perspective on randomness we take, it can be proven that *almost all* images of the entire set of possible images are *random.* Natural images, being clearly non-random, can thus only constitute a subset of vanishing size (Fig. 5).

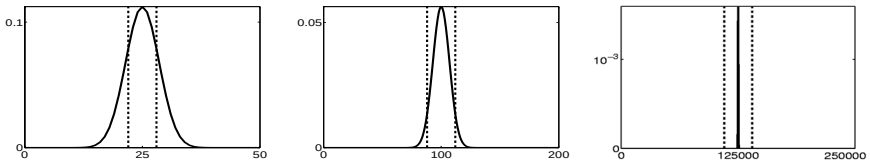


Fig. 4. The typical binary sequences can be defined by a constraint on the admissible deviation of the empirical distribution of 0's and 1's from equidistribution. The plots show the distribution of the number of different sequences in dependence of the number of 1's they contain. Sequence length is 50, 200, and 250.000 (the number of pixels in a typical image). The vertical lines illustrate a possible criterion for typicality (here about 50%+/-5%). It is obvious that for long sequences the criterion can be made arbitrarily tight, and almost all sequences will be typical

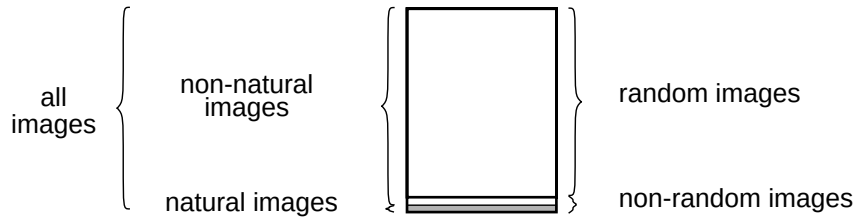


Fig. 5. Almost all possible images are random images

The observed breakdown of perceptual discriminatory performance for random images has thus the clear implication that *we are "functionally blind" for almost all possible images*. Stated this way, the seemingly unimportant perceptual equivalence of all random images should rather be seen as strong evidence for the adaptation hypothesis. Obviously, our visual system is not at all suited for an equally good processing of all kinds of images, but is rather highly specialized for the efficient processing of a very tiny subset: the subset that contains the natural images. This is certainly strong evidence in favor of the adaptation hypothesis.

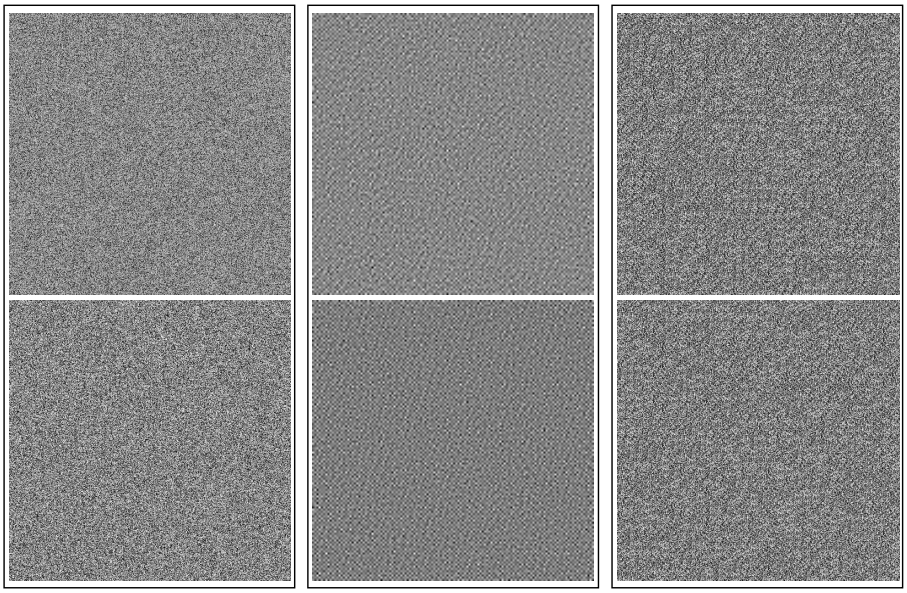
However, there exists an interesting generalization of the adaptation hypothesis which could also be consistent with the above observations. The crucial factor might not be seen in the *specific* regularities of the natural environment but rather in the mere fact that there *are* substantial statistical redundancies in the natural environment. The visual system may thus be specialized for specific images, but in a more general sense: it may work as a *universal structure detection system*.

How can we test this generic variant of the adaptation hypothesis? If the system would really be a universal structure detection system, then it should be able to distinguish between two images from any set which contains the same or less *amount* of statistical regularities as the set of natural images, but a different *type* of regularities. We hence considered different possibilities of constructing such quantitatively equivalent random processes. Furthermore, we constructed an artificial test set that is so simple that any reasonable universal structure detection system should be able to process it. Examples for these tests are shown in Figure 6. In all cases we get the same basic result: the images in these artificial test sets are almost impossible to distinguish.

Actually, a further related test is provided by what we usually employ as “random” numbers. In comparison to the typical size of an image any random number generator does only represent a very short description. Thus the “random” images it produces are certainly not really random. Nevertheless, they all appear visually as random, and cannot be distinguished. (By the way, the “random” images of Fig. 1 are simply two different subsequences of digits of π.) The generic version of the adaptation hypothesis, which assumes that the visual system could be a universal structure detection system, is therefore definitely falsified.

A last point to mention concerns the obviously limited complexity of all realistic information processing systems, which includes the visual system. It might be argued that, by definition, all those systems cannot adequately deal with signals of nearly unlimited complexity, like random images, and that the inability to discriminate such images is hence a trivial result. This conception is misleading, however. In order to

discriminate two random images it is not necessary to capture their full information content, but it is entirely sufficient to have a crude, low-complexity representation (as long as this representation is different for the two images). This is the case for a multi -



(a) (b) (c)

Fig. 6. Test of the hypothesis that the visual system is a universal structure detection system. In terms of the multivariate pdf $p(\vec{x})$ a simple equivalent class to the class of natural images is given by $p(\vec{y}) = p(A\vec{x})$, where A can be any orthonormal transform. Since such transforms represent mere rotations and reflections in state space, they leave the *shape* of the pdf, i.e., the *amount* of structure, and basic associated measures, like the information content (entropy), intact $(H(\vec{y}) = H(\vec{x}))$. (Of course, most k-order statistics measured with respect to the coordinate system will change). (a) A simple example are permutations. Clearly, a pdf $p(x_2,x_3,x_1)$ is in the above sense equivalent to $p(x_1,x_2,x_3)$. We can hence construct an artificial test set by a pseudo-random permutation (a low-complexity deterministic transform) of the state-space coordinates (i.e., the pixel positions) of natural images. Shown are the permuted versions of the two natural images of Fig. 1. (Since stationarity is destroyed, a strictly fair test would require to show several realizations to the system, but the other realizations look basically like the two shown). (b) If we consider stationarity crucial (the visual system may be a universal structure detection system only for stationary signals) we can first use a Fourier transform (an orthonormal transform), apply a permutation in the frequency domain, and perform then the inverse Fourier transform. Altogether, we obtain again an orthonormal transform, and the resulting signals should be stationary. We show again two examples from this artificial test set. (Note that the spectrum here is not white. The spectral permutation was pseudo-random, but with a constraint which avoids the occurrence of isolated high-amplitude peaks at low spatial frequencies.) (c) Finally, a critical test can be obtained by the provision of artificial images with extreme regularity (high redundancy). Each image of this set consists of 1024 patterns of size 16x16, randomly selected from an alphabet of 16 “basis” patterns. The information content is thus 512 Byte/image. The simple statistical structure can be easily detected by standard algorithms like KLT or Lempel-Ziv (each pattern is repeated about 64 times in each image). Of course, any universal structure detection system should be able to recognize the structure of such simple signals

channel, wavelet-like filter decomposition, the standard model of the visual system, and closely related to current image coding schemes (note that these have their quantization rules adapted to the statistics of typical images). Put random images into JPEG, and the probability that the coded images do not differ tends to zero.

3 Conclusion

Recent investigations indicate that the visual system is adapted to the statistical regularities of natural images. Here we examined whether this results in a behavioral difference in the perception of natural vs. non-natural images. First, we reconsidered the well known fact that human observers cannot distinguish random images, whereas they can easily distinguish natural images. We then asked how many images from the state space of possible images are random, and how many are natural. The answer from both a probabilistic perspective and from algorithmic complexity theory is that almost all images are random. Discrimination being a basic prerequisite for higher-level visual functions, the clear implication is that the visual system is functionally "blind" to almost all possible images. However, it works obviously quite well for the vanishingly small subset of natural images. This specialization cannot be attributed to a universal structure detection strategy, but seems to be crucially dependent on the "naturalness" of the structural constraints, since discrimination fails also for several non-random test images with non-natural statistical redundancies, even for very simple ones. Likewise, the discrimination failure cannot be attributed to a general complexity-constraint, since basic coding systems, like wavelet coders, yield clearly different representations for different random images. Together, these results can be regarded as strong behavioral evidence for the hypothesis that the processing structures of early vision are the result of an optimized adaptation to the specific statistical redundancies of natural images.

Acknowledgment: Study supported by DFG (SFB 462, GRK 267). I thank U. Nuding for intense discussions, and G. Hauske and I. Rentschler for helpful comments.

References

1. C. Zetzsche and G. Krieger. Nonlinear mechanisms and higher-order statistics in biological vision and electronic image processing: review and perspectives. J. Electronic Imaging, **10**(1) (2001) 56–99
2. D. Tolhurst and Y. Tadmor. Band-limited contrast in natural images explains the detectability of changes in the amplitude spectra. Vision Research, **37**(23) (1997) 3203–3215
3. D. Tolhurst and Y. Tadmor. Discrimination of spectrally blended natural images: optimisation of the human visual system for encoding natural images. Perception, **29**(12) (2000) 1087–1100
4. P. Martin-Löf. The definition of random sequences. Inform. and Control **9** (1966) 602–619
5. G. Chaitin. *Algorithmic Information Theory.* Cambridge Univ. Press, Cambridge (1987)
6. M. Li and P. Vitanyi. *An introduction to Kolomogorov complexity and its applications*. Springer, New York, Berlin (1993)

Recognition and matching

Face detection by using independent component decomposition

I. Buciu C. Kotropoulos I. Pitas *

Department of Informatics, Aristotle University of Thessaloniki
GR-540 06, Thessaloniki, Box 451, Greece, {costas,pitas}@zeus.csd.auth.gr

Abstract. In this paper we explore the independent component decomposition for face detection. The minimization of the Kullback - Leibler divergence and the maximization of the entropy are two methods employed to decompose an original image into its independent components. We built nearest neighbor classifiers based on their resulting independent components and compare their ability to detect faces to that of support vector machines.

1 Introduction

There are many applications in which human face detection plays a very important role. For example, it can be used in content-based image database indexing/searching, surveillance systems, and human-centered computer interfaces. It also constitutes the first step in a fully automatic face recognition system. A comprehensive survey on face detection methods is given in [1]. A face detection technique based on independent component decomposition is developed in this paper. The principal components matrix of the original face and non-face patterns is assumed to represent a mixture of independent image sources which are retrieved by using independent component analysis (ICA) through an unmixing matrix. We can reconstruct the original images by combining linearly these sources. The matrix which contains the coefficients of those combinations is further use as the first input of the two nearest neighbor classifiers employed in the paper. The second input is a combination of the test image with principal components matrix and the unmixing matrix. The classification is then performed according to the nearest neighbor rule. Testing this approach against support vector machines (SVMs), we found the latter is outperformed by the proposed method in the face detection task.

2 Spatial independent component analysis

The goal of is to decompose a set of observations into a basis whose components are statistically independent or, at least, are as independent as possible. ICA

* This work was supported by the European Union Research Training Network "Multimodal Human-Computer Interaction (HPRN-CT-2000-00111).

originally applied to blind source separation [2]. Two ICA representations of facial patterns have been proposed in [3] for face recognition. The discriminating ability of ICA alone or when combined with other discriminant criteria, such as Bayesian framework or Fisher's linear discriminant, was analyzed in [4].

In our analysis we follow the model proposed in [3]. Consider a matrix $\mathbf{X}$ whose rows contain vectors formed by scanning lexicographically face and non-face patterns (i.e., image regions). We assume that $\mathbf{X}$ contains a mixture of the original independent sources $\mathbf{U}$. The matrix is decomposed into a family of $\mathbf{Y}$ independent sources passing it through an unmixing matrix $\mathbf{D}$ in the attempt to recover $\mathbf{U}$. Each source (row of $\mathbf{Y}$) is an image whose pixel values are independent of those in every other image. Accordingly, these images are said to be spatially independent. We refer to this model as the *spatial* ICA. Having a number of n face and non-face images, the number of independent components will be n as well. In order to have a control on the number of independent components, we choose m linear combinations of face and non-face patterns, namely the principal component vectors of the image set. Let $\mathbf{P}_m^T$ denote the matrix that is formed by the m principal components in its rows. The objective of ICA applied onto $\mathbf{P}_m^T$ is to find the matrix $\mathbf{Y}$ whose rows are the statistically independent sources by appropriately determining the unmixing matrix $\mathbf{D}$. The relationship between the three aforementioned matrices is given by [3]:

$$\mathbf{Y} = \mathbf{D}\mathbf{P}_m^T. \tag{1}$$

Frequently, a *whitening* process applied to $\mathbf{P}_m^T$ is necessary to decorrelate and normalize the data. If the row means are substracted from $\mathbf{P}_m^T$ and the resulting matrix is passed through a zero-phase whitening filter which is twice the inverse square root, the whitening transformation is written as $\mathbf{W} = 2(\mathbf{P}_m^T\mathbf{P}_m)^{-\frac{1}{2}}$. Therefore, the zero - mean input matrix can be computed as the product of the unmixing matrix and the whitening matrix $\mathbf{D}_w = \mathbf{D}\mathbf{W}$. Eq. (1) is rewritten as follows:

$$\mathbf{Y} = \mathbf{D}_w\mathbf{P}_m^T \Longrightarrow \mathbf{P}_m^T = \mathbf{D}_w^{-1}\mathbf{Y}. \tag{2}$$

The reconstructed image by ICA is:

$$\mathbf{X}_{recICA} = (\mathbf{X}\mathbf{P}_m\mathbf{D}_w^{-1})\mathbf{Y} = \mathbf{C}_{train}\mathbf{Y}. \tag{3}$$

The matrix $\mathbf{C}_{train}$ contains the coefficients of the linear combination of spatial independent sources $\mathbf{Y}$. Each row of $\mathbf{Y}$ comprises the independent component representation of the face images. Once we have finished training and obtained $\mathbf{Y}$, a test image can be presented as:

$$\mathbf{c}_{test} = \mathbf{D}_w^{-1}\mathbf{P}_m\mathbf{x}_{test}. \tag{4}$$

2.1 Entropy maximization

Given $\mathbf{P}_m^T$, the component in (1) which is responsible for obtaining the independent sources is the unmixing matrix $\mathbf{D}$ that must be updated in order to

obtain sources that are as independent as possible. Different approaches exist for this purpose. One way is the so called maximum entropy method which has been developed in [5]. The matrix $\mathbf{Y}$ is transformed into a matrix $\mathbf{Z}$ by passing it through a component-wise nonlinearity denoted by $\mathbf{G}[\cdot]$. As ICA is applied on the columns of $\mathbf{P}_m^T$, a realization $\mathbf{p}_j$ is a combination of the original sources $\mathbf{u}_j$ via a mixing matrix $\mathbf{A}$, $\mathbf{p}_j = \mathbf{A}\mathbf{u}_j$. Therefore, the sources can be restored through the unmixing matrix $\mathbf{D}$ as $\mathbf{y}_j = \mathbf{D}\mathbf{p}_j \approx \mathbf{u}_j$. For simplicity we omit the index j from now on. Passing the sources $\mathbf{y}$ through $\mathbf{G}$ yields:

$$\mathbf{z} = \mathbf{G}(\mathbf{y}) = \mathbf{G}(\mathbf{D}\mathbf{p}) = \mathbf{G}(\mathbf{D}\mathbf{A}\mathbf{u}). \tag{5}$$

Therefore:

$$\mathbf{u} = \mathbf{A}^{-1}\mathbf{D}^{-1}\mathbf{G}^{-1}(\mathbf{z}) = \mathbf{\Psi}(\mathbf{z}). \tag{6}$$

The entropy is given by:

$$h(\mathbf{z}) = -E[\log(f_{\mathbf{Z}}(z))] = -E\left[\log\left(\frac{f_{\mathbf{U}}(\mathbf{u})}{|\det(\mathbf{J}(\mathbf{u}))|}\right)\right], \tag{7}$$

where $f_{\mathbf{Z}}(\mathbf{z})$ and $f_{\mathbf{U}}(\mathbf{u})$ are the probability density functions of $\mathbf{Z}$ and the sources $\mathbf{U}$, and $\mathbf{J}$ is the Jacobian matrix $\mathbf{J} = \partial\mathbf{z}/\partial\mathbf{y}$. Using the chain rule, the determinant of $\mathbf{J}$ can be evaluated as:

$$|\det(\mathbf{J}(\mathbf{u}))| = \left|\det\left(\frac{\partial\mathbf{z}}{\partial\mathbf{y}}\right)\right| = |\det(\mathbf{D}\mathbf{A})| \prod_{i=1}^{m} \frac{\partial z_i}{\partial y_i}. \tag{8}$$

Maximizing the entropy $h(\mathbf{z})$ requires to maximize the expectation of the denominator term $\log|\det(\mathbf{J}(\mathbf{u}))|$ with respect to the matrix $\mathbf{D}$:

$$\frac{\partial}{\partial\mathbf{D}}(\log|\det(\mathbf{J}(\mathbf{u}))|) = [\mathbf{D}^{-1}]^T + \sum_{i=1}^{m} \frac{\partial}{\partial\mathbf{D}} \log\left(\frac{\partial z_i}{\partial y_i}\right). \tag{9}$$

If $z_i = g(y_i) = 1/(1+e^{-y_i})$ is a component-wise nonlinearity applied to all elements of matrix $\mathbf{Y}$, and taking into account that:

$$\frac{\partial z_i}{\partial s_i} = z_i(1-z_i), \tag{10}$$

and $\mathbf{y} = \mathbf{G}^{-1}(\mathbf{z})$, (9) becomes:

$$\frac{\partial}{\partial\mathbf{D}}(\log|\det(\mathbf{J}(\mathbf{s}))|) = [\mathbf{D}^{-1}]^T + (\mathbf{1}-2\mathbf{z})\mathbf{p}^T. \tag{11}$$

Using the gradient ascent algorithm, the change of the unmixing matrix $\mathbf{D}$ is [5]:

$$\Delta\mathbf{D} = \eta(\mathbf{D}^{-T} + (\mathbf{1}-2\mathbf{z})\mathbf{p}^T). \tag{12}$$

It is more convenient to use the natural gradient instead of the actual one to avoid inverting $\mathbf{D}$ at each step, therefore, the formula for unmixing matrix change becomes:

$$\mathbf{D}_{k+1} = \eta[\mathbf{I} + (\mathbf{1}-2\mathbf{z})\mathbf{y}^T]\mathbf{D}_k. \tag{13}$$

2.2 Minimization of the Kullback-Leibler divergence

Another way to obtain independent sources is equivalent with minimizing the Kullback-Leibler divergence between the probability density function $f_{\mathbf{S}}(\mathbf{s};\mathbf{D})$ parameterized by $\mathbf{D}$ and the corresponding factorial distribution defined by [6]:

$$\widehat{f_{\mathbf{Y}}}(\mathbf{y};\mathbf{D}) = \prod_{i=1}^{m} \widehat{f_{\mathbf{Y}}}(\mathbf{y}_i;\mathbf{D}). \tag{14}$$

The Kullback-Leibler divergence is given by:

$$\mathcal{D}_{f\|\widehat{f}}(\mathbf{D}) = -h(\mathbf{y}) + \sum_{i=1}^{m} \widehat{h}(\mathbf{y}_i), \tag{15}$$

where $h(\mathbf{y})$ is the entropy of the random vector $\mathbf{y}$ at the output of the unmixer and $\widehat{h}(\mathbf{y}_i)$ is the marginal entropy of the ith element of $\mathbf{y}$. The minimization can be implemented using the method of gradient descent. Following [6], the unmiximg matrix will be updated at each iteration k as follows:

$$\mathbf{D}_{k+1} = \mathbf{D}_k + \eta[\mathbf{I} - \theta(\mathbf{y}_k)\mathbf{y}_k^T]\mathbf{D}_k^{-T}, \tag{16}$$

where $\mathbf{I}$ is the identity matrix and the analytical form of the *activation function* $\theta(\mathbf{y})$ is also given by [6].

3 ICA performance evaluation

The ability of ICA for face detection was evaluated using face patterns derived from the AT&T face database. A description of the data is given in [7]. A number of 294 non-face patterns was collected and added to 306 face patterns, achieving a total data base of 600 patterns. 80 of them were used to form the training set. Each row of the training matrix contains a 238 - dimensional vector. This matrix was updated according to (13) and (16) for the first and second method respectively, for 1000 iterations. The learning rate η was set to 10^{-6}. The evaluation of the ICA performance was assessed by means of two classifiers. The first one is based on the nearest neighbor rule and measures the angle between a test vector and a training one. Let us denote the class of face feature vectors by $\mathcal{L}_1$ and those of the non-face feature vectors by $\mathcal{L}_{-1}$. Let $\mathbf{c}_{+1}$ be a row vector of $\mathbf{C}_{train}$ matrix that corresponds to the nearest face pattern. Let us denote the nearest non-face neighbor of $\mathbf{c}_{test}$ by $\mathbf{c}_{-1}$. Then we compute the quantities:

$$d_f = \frac{\mathbf{c}_{test}^T\mathbf{c}_{+1}}{\|\mathbf{c}_{test}\|\|\mathbf{c}_{+1}\|} \quad \text{and} \quad d_{nf} = \frac{\mathbf{c}_{test}^T\mathbf{c}_{-1}}{\|\mathbf{c}_{test}\|\|\mathbf{c}_{-1}\|}, \tag{17}$$

where d_f and d_{nf} are the cosines of the angle between a test feature vector and the nearest training one. We assign $\mathbf{c}_{test}$ to $\mathcal{L}_1$ if $d_f > d_{nf}$, otherwise $\mathbf{c}_{test} \in \mathcal{L}_{-1}$. Notice that the labels for the training set are preserved, therefore we know the labels corresponding to $\mathbf{C}_{train}$.

The second classifier is the a minimum Euclidean distance classifier. The Euclidean distance from $\mathbf{c}_{test}$ to $\mathbf{c}_k$, where $k \in \{\pm 1\}$ is expressed as

$$\begin{aligned}\|\mathbf{c}_{test} - \mathbf{c}_k\|^2 &= -2[\mathbf{c}_k^T \mathbf{c}_{test} - \frac{1}{2}\mathbf{c}_k^T \mathbf{c}_k] + \mathbf{c}_{test}^T \mathbf{c}_{test} \\ &= -2h_k(\mathbf{c}_{test}) + \mathbf{c}_{test}^T \mathbf{c}_{test}, \end{aligned} \tag{18}$$

where $h_k(\mathbf{c}_{test})$ is a linear discriminant function of $\mathbf{c}_{test}$. A test pattern is classified by this classifier (also known as "maximum correlation classifier") by computing two linear discriminant function $h_{+1}(\mathbf{c}_{test})$ and $h_{-1}(\mathbf{c}_{test})$ and assigning $\mathbf{c}_{test}$ to the class corresponding to the maximum discriminant function.

We have investigated the performance of the two previously mentioned classifiers (17) and (18) by varying the number of principal components extracted from the training set. The results are depicted in Figure 1. A minimum error of 5.2% was achieved using 20 principal components in the case of the second classifier. However, the performance of this classifier seems to be almost insensitive to the number of the principal components used. On the contrary, for the nearest neighbor rule, the classification error decreases as the number of principal components involved increases. A minimum 3.9% classification error is achieved by keeping 70 linear combinations of 80 training vectors. For comparison, support vector machines (SVMs) with different kernels [8] were applied to discriminate between the face and the non-face patterns. The error rates for different SVMs are included the Table 1, in the same experiment for comparison purposes.

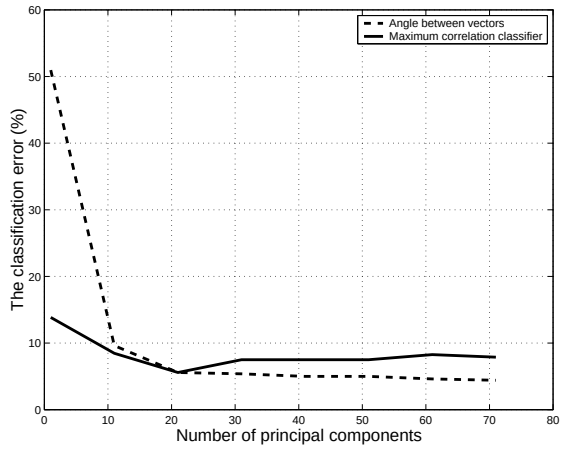


Fig. 1. Classification error (false acceptance rate plus false rejection rate) versus the number of principal components for both classifiers.

Table 1. Number of errors (%) for several classifiers.

Face detection methods	Errors (%)
ICA-based classifier 1	3.9
ICA-based classifier 2	5.2
linear SVM	6.1
polynomial SVM with degree equals 2	6.3
polynomial SVM with degree equals 3	11.1
radial basis function SVM	5.5
exponential radial basis function SVM	6.1

4 Conclusions

We have exploited the ability of ICA to provide useful features in order to conduct a face detection task. The combination of ICA with nearest neighbor classifiers seems to provide a reliable face detector that can outperform SVMs.

References

1. M.-H. Yang, N. Ahuja, and D. Kriegman, "A survey on face detection methods," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, January 2002.
2. J.F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE, Special Issue on Blind Identification and Estimation*,vol. 90, no. 8, pp. 2009–2026, October, 1998.
3. M.S. Bartlett, H.M. Lades, and T.J. Sejnowski, "Independent component representations for face recognition," in *Proc. SPIE Conf. on Human Vision and Electronic Imaging III*, vol. 3299 pp. 528–539, 1998.
4. C. Liu and H. Wechsler, "Comparative Assessment of Independent Component Analysis (ICA) for Face Recognition," in *Proc. Second Int. Conf. on Audio- and Video-based Biometric Person Authentication*, 1999.
5. A.J. Bell and T.J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 6, pp. 1129–1159, 1995.
6. S. Haykin, *Neural Networks. A Comprehensive Foundation*. New Jersey: Prentice-Hall, Inc. 1999.
7. I. Buciu, C. Kotropoulos and I. Pitas, "Combining support vector machines for accurate face detection," in *Proc. 2001 IEEE Int. Conf. on Image Processing*, pp 1054–1057, 2001.
8. V.N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.

A General Model for the Development of Retinotopic Projections between Manifolds of Different Geometries

M. Güßmann[1], A. Pelster[2], and G. Wunner[1]

[1]Institut für Theoretische Physik 1, Universität Stuttgart, 70550 Stuttgart, Germany
[2]Institut für Theoretische Physik, Freie Universität Berlin, 14195 Berlin, Germany

Abstract. We generalize a one-dimensional model of Häussler and von der Malsburg which describes the generation of retinotopic projections between two cell sheets. Our generalized model is independent of the special geometry of the cell array and describes the temporal evolution of the connection strengths between cells on different manifolds. Linearizing the equations of evolution around the stationary homogeneous state and using of the methods of synergetics leads to order parameter equations near the instability which contain only the unstable modes. We show that our general model contains as a special case the description of cell sheets of Häussler and von der Malsburg, and that it allows a detailed treatment of cell arrays distributed on spherical shells.

1 Introduction

In the course of ontogenesis of vertebrate animals well-ordered neural connections are established between retina and tectum, a part of the brain which plays an important role in processing optical information. As a result of this selforganization process neighbouring retinal cells project onto neighbouring cells of the tectum. Such a projection is called *retinotopic*. This conservation of neighbourhood relations is realized in many neural connections between different sheets of cells.

A detailed analytical treatment of development of ordered projections between different sheets of nerve cells was already presented by Häussler and von der Malsburg [1]. In that paper retina and tectum were treated as one-dimensional discrete arrays of cells. The case of continuously distributed cells on a spherical shell was discussed partially in [2].

2 Our Model

Here we generalize the approach by developing a model which is *independent* of the special geometry of the problem. To that end retina and tectum are represented by general manifolds M_t and M_r, respectively. We define a measure

for the magnitudes of the manifolds by

$$|M_t| = \int\!\!\!\!\!\!\sum_{\mathbf{t}} 1\,, \qquad |M_r| = \int\!\!\!\!\!\!\sum_{\mathbf{r}} 1\,. \tag{1}$$

The symbol

$$\int\!\!\!\!\!\!\sum$$

stands for a summation over all elements of M_t, M_r, if the manifolds are discrete, and for an integration if the manifolds are continuous. As we restrict our investigations to manifolds of identical topology, we have

$$|M_t| = |M_r| := M\,. \tag{2}$$

Every ordered pair $(\mathbf{t}, \mathbf{r})$ with $\mathbf{t} \in M_t$, $\mathbf{r} \in M_r$ is connected by a *connection strength* $w(\mathbf{t}, \mathbf{r})$. The equations of evolution of these connections are assumed to be given by a generalization of the *Häussler equations* [1]:

$$\begin{aligned}\dot{w}(\mathbf{t},\mathbf{r}) = \alpha + w(\mathbf{t},\mathbf{r}) \int\!\!\!\!\!\!\sum_{\mathbf{t}'} \int\!\!\!\!\!\!\sum_{\mathbf{r}'} c_T(\mathbf{t},\mathbf{t}')c_R(\mathbf{r},\mathbf{r}')w(\mathbf{t}',\mathbf{r}') \\ - \frac{w(\mathbf{t},\mathbf{r})}{2M}\left[\int\!\!\!\!\!\!\sum_{\mathbf{t}'}\left\{\alpha + w(\mathbf{t}',\mathbf{r})\int\!\!\!\!\!\!\sum_{\mathbf{t}''}\int\!\!\!\!\!\!\sum_{\mathbf{r}'} c_T(\mathbf{t}',\mathbf{t}'')c_R(\mathbf{r},\mathbf{r}')w(\mathbf{t}'',\mathbf{r}')\right\}\right. \\ \left. + \int\!\!\!\!\!\!\sum_{\mathbf{r}'}\left\{\alpha + w(\mathbf{t},\mathbf{r}')\int\!\!\!\!\!\!\sum_{\mathbf{t}'}\int\!\!\!\!\!\!\sum_{\mathbf{r}''} c_T(\mathbf{t},\mathbf{t}')c_R(\mathbf{r}',\mathbf{r}'')w(\mathbf{t}',\mathbf{r}'')\right\}\right] .\end{aligned} \tag{3}$$

Here α denotes the homogeneous growth-rate of new synapses onto the tectum, and the positive coefficients $c_T(\mathbf{t}, \mathbf{t}')$, $c_R(\mathbf{r}, \mathbf{r}')$ represent measures for cooperativity within each manifold, which are larger when the points $\mathbf{t}$, $\mathbf{t}'$ and $\mathbf{r}$, $\mathbf{r}'$ are closer to each other, and fulfill the normalization condition

$$\int\!\!\!\!\!\!\sum_{\mathbf{t}'} c_T(\mathbf{t},\mathbf{t}') = 1\,, \qquad \int\!\!\!\!\!\!\sum_{\mathbf{r}'} c_R(\mathbf{r},\mathbf{r}') = 1\,. \tag{4}$$

3 Orthonormal System

Furthermore we assume spatial homogeneity and isotropy of the manifolds, i.e., no point is preferred to another, and no direction is preferred to another. As mentioned above the strength of cooperation depends on the distance between two points of the manifold. This requires a measure of distance, i.e. a metric, which turns out to be the stationary *Robertson–Walker metric* of general relativity [3]. With the help of the metric we define the Laplace–Beltrami operators Δ_{M_t}, Δ_{M_r} on the manifolds and use their eigenvalue problems

$$\Delta_{M_t}\psi^{m_T}_{\lambda_T}(\mathbf{t}) = \lambda_T \psi^{m_T}_{\lambda_T}(\mathbf{t})\,, \tag{5}$$

$$\Delta_{M_r}\psi^{m_R}_{\lambda_R}(\mathbf{r}) = \lambda_R \psi^{m_R}_{\lambda_R}(\mathbf{r})\,, \tag{6}$$

to define a complete orthonormal system with the eigenfunctions $\psi_{\lambda_T}^{m_T}(\mathbf{t})$ and $\psi_{\lambda_R}^{m_R}(\mathbf{r})$. By construction, they fulfill the orthonormality relations

$$\sum_{\mathbf{t}} \psi_{\lambda_T}^{m_T}(\mathbf{t})\psi_{\lambda'_T}^{m'_T *}(\mathbf{t}) = \delta_{\lambda_T\lambda'_T}\delta_{m_T m'_T}\,, \tag{7}$$

$$\sum_{\mathbf{r}} \psi_{\lambda_R}^{m_R}(\mathbf{r})\psi_{\lambda'_R}^{m'_R *}(\mathbf{r}) = \delta_{\lambda_R\lambda'_R}\delta_{m_R m'_R}\,, \tag{8}$$

and the completeness relations

$$\sum_{\lambda_T}\sum_{m_T} \psi_{\lambda_T}^{m_T}(\mathbf{t})\psi_{\lambda_T}^{m_T *}(\mathbf{t}') = \delta(\mathbf{t}-\mathbf{t}')\,, \tag{9}$$

$$\sum_{\lambda_R}\sum_{m_R} \psi_{\lambda_R}^{m_R}(\mathbf{r})\psi_{\lambda_R}^{m_R *}(\mathbf{r}') = \delta(\mathbf{r}-\mathbf{r}')\,. \tag{10}$$

Here the indices m_T, m_R denote the degeneracy of the eigenspaces belonging to the eigenvalues λ_T, λ_R. The cooperativity coefficients can be expanded in terms of these functions according to

$$c_T(\mathbf{t},\mathbf{t}') = \sum_{\lambda_T}\sum_{\lambda'_T}\sum_{m_T}\sum_{m'_T} F_{\lambda_T\lambda'_T}^{m_T m'_T}\psi_{\lambda_T}^{m_T}(\mathbf{t})\psi_{\lambda'_T}^{m'_T *}(\mathbf{t}')\,, \tag{11}$$

$$c_R(\mathbf{r},\mathbf{r}') = \sum_{\lambda_R}\sum_{\lambda'_R}\sum_{m_R}\sum_{m'_R} F_{\lambda_R\lambda'_R}^{m_R m'_R}\psi_{\lambda_R}^{m_R}(\mathbf{r})\psi_{\lambda'_R}^{m'_R *}(\mathbf{r}')\,, \tag{12}$$

where we assume for the sake of simplicity

$$F_{\lambda_T\lambda'_T}^{m_T m'_T} = f_{\lambda_T}^{m_T}\delta_{\lambda_T\lambda'_T}\delta_{m_T m'_T}\,, \tag{13}$$

$$F_{\lambda_R\lambda'_R}^{m_R m'_R} = f_{\lambda_R}^{m_R}\delta_{\lambda_R\lambda'_R}\delta_{m_R m'_R}\,, \tag{14}$$

with expansion coefficients $f_{\lambda_T}^{m_T}$, $f_{\lambda_R}^{m_R}$. The forms (13), (14) are essential assumptions, and thus ingredients, of our description. We then have

$$c_T(\mathbf{t},\mathbf{t}') = \sum_{\lambda_T}\sum_{m_T} f_{\lambda_T}^{m_T}\psi_{\lambda_T}^{m_T}(\mathbf{t})\psi_{\lambda_T}^{m_T *}(\mathbf{t}')\,, \tag{15}$$

$$c_R(\mathbf{r},\mathbf{r}') = \sum_{\lambda_R}\sum_{m_R} f_{\lambda_R}^{m_R}\psi_{\lambda_R}^{m_R}(\mathbf{r})\psi_{\lambda_R}^{m_R *}(\mathbf{r}')\,. \tag{16}$$

Note that the normalization of the cooperation coefficients (4) fixes

$$f_0^{m_T} = f_0^{m_R} = 1\,. \tag{17}$$

4 Linear Stability Analysis

Using the methods of synergetics [4, 5] the system is investigated around the stationary homogeneous solution $w(\mathbf{t}, \mathbf{r}) = 1$. A linearization of the generalized Häussler equations (3) with respect to the deviation

$$v(\mathbf{t}, \mathbf{r}) = w(\mathbf{t}, \mathbf{r}) - 1 \tag{18}$$

leads to the eigenvalue problem

$$L(\mathbf{t}, \mathbf{r}, v(\mathbf{t}, \mathbf{r})) = \Lambda\, v(\mathbf{t}, \mathbf{r}) \tag{19}$$

with a linear operator L. The eigenfunctions $v(\mathbf{t}, \mathbf{r})$ turn out to be

$$v(\mathbf{t}, \mathbf{r}) = \psi_{\lambda_T}^{m_T}(\mathbf{t}) \psi_{\lambda_R}^{m_R}(\mathbf{r})\,, \tag{20}$$

with the eigenvalues Λ given by

$$\Lambda_{\lambda_T \lambda_R}^{m_T m_R} = \begin{cases} -\alpha - 1 & \lambda_T = \lambda_R = 0 \\ -\alpha + \frac{1}{2}(f_{\lambda_T}^{m_T} f_{\lambda_R}^{m_R} - 1) & \lambda_T = 0, \lambda_R \neq 0; \lambda_R = 0, \lambda_T \neq 0 \\ -\alpha + f_{\lambda_T}^{m_T} f_{\lambda_R}^{m_R} & \text{otherwise}\ . \end{cases} \tag{21}$$

5 Nonlinear Analysis

By changing the control parameter α in the generalized Häussler equations (3) in a suitable way the real parts of some eigenvalues become positive, therefore the system can be driven to the neighbourhood of an instability point. If we assume that the expansion coefficients $f_{\lambda_T}^{m_T}$, $f_{\lambda_R}^{m_R}$ are monotonous with respect to λ_T, λ_R,

$$1 = f_0^{m_T} \geq f_1^{m_T} \geq f_2^{m_T} \geq \cdots \geq 0\,, \tag{22}$$

$$1 = f_0^{m_R} \geq f_1^{m_R} \geq f_2^{m_R} \geq \cdots \geq 0\,, \tag{23}$$

then the maximum eigenvalue is given by $\Lambda_{11}^{m_T m_R}$.The linear stability analysis reveals in (21) that the neighbourhood of the instability point is characterized by

$$\Re(\Lambda_{11}^{m_T m_R}) \approx 0\,, \tag{24}$$

$$\Re(\Lambda_{\lambda_T \lambda_R}^{m_T m_R}) < 0\,, \quad (\lambda_T, \lambda_R) \neq (1, 1)\,. \tag{25}$$

Thus the amounts of the real parts of the unstable modes $(\lambda_T, \lambda_R) = (1, 1)$ are much smaller than those of the stable modes $(\lambda_T, \lambda_R) \neq (1, 1)$:

$$|\Re(\Lambda_{11}^{m_T m_R})| \ll |\Re(\Lambda_{\lambda_T \lambda_R}^{m_T m_R})|\,, \qquad (\lambda_T, \lambda_R) \neq (1, 1)\,. \tag{26}$$

This result motivates decomposing the connection strength according to

$$w(\mathbf{t}, \mathbf{r}, t) = 1 + U(\mathbf{t}, \mathbf{r}, t) + S(\mathbf{t}, \mathbf{r}, t)\,, \tag{27}$$

where

$$U(\mathbf{t},\mathbf{r},t) = \sum_{m_T m_R} U_{11}^{m_T m_R}(t)\psi_1^{m_T}(\mathbf{t})\psi_1^{m_R}(\mathbf{r}) \tag{28}$$

denotes the contribution of the unstable modes and

$$S(\mathbf{t},\mathbf{r},t) = \sum\!\!\!\!\!\!\int_{\lambda_T\lambda_R}' \sum_{m_T m_R} S_{\lambda_T\lambda_R}^{m_T m_R}(t)\psi_{\lambda_T}^{m_T}(\mathbf{t})\psi_{\lambda_R}^{m_R}(\mathbf{r}) \tag{29}$$

denotes the contribution of the stable modes. The symbol

$$\sum\!\!\!\!\!\!\int_{\lambda_T\lambda_R}'$$

means the summation/integration over all eigenvalues λ_T and λ_R except for $(\lambda_T,\lambda_R) = (1,1)$.

Relation (26) leads to the *time-scale hierarchy*, i.e. the stable modes evolve on a faster time-scale than the unstable modes,

$$\tau_u = [\Re(\Lambda_{11}^{m_T m_R})]^{-1} \gg \tau_s = [\Re(\Lambda_{\lambda_T\lambda_R}^{m_T m_R})]^{-1}\,. \tag{30}$$

Due to this time-scale hierarchy the dynamics of the stable modes quasi-instantaneously follow the dynamics of the unstable modes:

$$S(\mathbf{t},\mathbf{r},t) = h(U(\mathbf{t},\mathbf{r},t))\,. \tag{31}$$

This is the well-known *slaving principle* of synergetics: the stable modes are enslaved by the unstable modes. The *center manifold* $h(U(\mathbf{t},\mathbf{r},t))$ is calculated by eliminating the stable modes. Thus it is possible to reduce the original high-dimensional system to a low-dimensional one which only contains the unstable amplitudes. The general form of these order parameter equations is independent of the geometry of the problem and reads

$$\dot{U}_{11}^{m_T m_R}(t) = \Lambda_{11}^{m_T m_R} U_{11}^{m_T m_R}(t) + \sum_{m_T' m_R'} \sum_{m_T'' m_R''} A_{m_T m_T' m_T''}^{m_R m_R' m_R''} U_{11}^{m_T' m_R'}(t) U_{11}^{m_T'' m_R''}(t)$$

$$+ \sum_{m_T' m_R'} \sum_{m_T'' m_R''} \sum_{m_T''' m_R'''} B_{m_T m_T' m_T'' m_T'''}^{m_R m_R' m_R'' m_R'''} U_{11}^{m_T' m_R'}(t) U_{11}^{m_T'' m_R''}(t) U_{11}^{m_T''' m_R'''}(t)\,, \tag{32}$$

where the coefficients $A_{m_T m_T' m_T''}^{m_R m_R' m_R''}$, $B_{m_T m_T' m_T'' m_T'''}^{m_R m_R' m_R'' m_R'''}$ can be expressed in terms of the eigenfunctions $\psi_{\lambda_T}^{m_T}(\mathbf{t})$, $\psi_{\lambda_R}^{m_R}(\mathbf{r})$ and the expansion coefficients $f_{\lambda_T}^{m_T}$, $f_{\lambda_R}^{m_R}$ [6]. As is usual in synergetics, the coefficients $B_{m_T m_T' m_T'' m_T'''}^{m_R m_R' m_R'' m_R'''}$ in general consist of two parts, one stemming from the order parameters themselves and the other representing the influence of the center manifold. Equations (32) are a central new result of this paper, and serve as a starting point of the analysis of selforganization in cell arrays of different geometries.

6 Examples

Specifying the geometry means inserting the corresponding eigenfunctions of the Laplacian into the order parameter equations (32). For the linear chain these eigenfunctions are given by periodic exponential functions, and we regain the results presented in [1]. For the case of spherical shells the eigenfunctions are given by spherical harmonics:

$$Y_{lm}(\vartheta, \varphi)\,, \qquad l = 0, 1, 2, \ldots; \quad m = -l, -l+1, \ldots, l-1, l\,. \tag{33}$$

The calculation of the order parameter equations (32) for the spherical shell shows that the quadratic term vanishes, by analogy with the linear chain. The cubic part contains only terms which fulfill the conditions

$$m'_T + m''_T + m'''_T = m_T\,, \tag{34}$$

$$m'_R + m''_R + m'''_R = m_R\,. \tag{35}$$

It turns out that the dynamics of the order parameters for the spherical shell can be described by a *potential* V, which was also the case for the linear chain [1]. Because the coefficients of (32) are quite complicated expressions we have to refer the reader to [6] for a detailed presentation of the order parameter equations and the potential V.

7 Outlook

The Robertson–Walker metric describes manifolds with constant curvature. In [6] we revisit the linear chain, which represents a Euclidean manifold with curvature 0, and we treat the spherical shell, which represents a curved manifold with curvature +1. There remains the interesting task of investigating the case of a *non-Euclidean* manifold with negative curvature, namely the *pseudosphere* (curvature –1) [3].

References

1. Häussler, A.F. and von der Malsburg, C.: *Development of Retinotopic Projections: An Analytical Treatment.* J. Theoret. Neurobiol. **2**, 47 (1983)
2. Wagner, W. and von der Malsburg, C.: unpublished result
3. Wunner, G., Güßmann, M., and Wunderlin, A.: *Selforganization between Manifolds with $d > 1$ in Euclidean and non–Euclidean Geometry by Cooperation and Competition.* Talk at the University of Bochum (May 2001)
4. Haken, H.: *Synergetics, An Introduction.* Springer, Berlin, 3rd ed. (1983)
5. Haken, H.: *Advanced Synergetics.* Springer, Berlin (1983)
6. Güßmann, M., Pelster, A., and Wunner, G.: in preparation

Dynamic Link Matching of Severely Deformed Patterns by General Local Linear Maps

Florian Hardt and Günter Wunner

Institut für Theoretische Physik 1, Universität Stuttgart, 70550 Stuttgart, Germany

Abstract. The extension of dynamic link matching by introducing local linear maps (LLMs) has been proposed to render the matching adaptable to larger deformations. However, investigations in the literature so far have been restricted to local rotations, i.e. the Jacobian J of the map is a simple 2-d rotation matrix. Here we will describe the generalization of this approach. We make use of the theorem that every Jacobian in two dimensions can be decomposed into a rotation by some angle γ_1, followed by stretchings λ_1, λ_2 in both directions, and another rotation by an angle γ_2. While in the previous model only one parameter for the rotation was needed, the general LLM has to include the full set of parameters γ_1, γ_2, λ_1, λ_2. The decomposition allows for a natural classification of LLMs in generic subclasses. As an example of a generic two-parameter map we discuss conformal local linear maps.

1 Introduction

Dynamic Link Matching (DLM) (e.g. [1]) is a well-known pattern matching model tolerant of small deformations. Local features are extracted from the data pattern and are matched to similar counterparts in a template pattern. The matching process fails when due to strong deformations extracted features of corresponding points are no longer similar. To increase the model's tolerance to deformations the introduction of Local Linear Maps (LLMs) was suggested [2]. LLMs approximate the local deformation of the data pattern. By this, extracted features become invariant under all deformations within the range of the LLM applied. The LLM modifies the filters used for feature extraction in the same manner as the data pattern is distorted. Thus appropriately transformed Gabor type filters can extract features in the data pattern similar to those in the template pattern. Moreover, LLMs of neighbouring data points are required to have smoothly varying characteristics, which enforces topological constraints. (In a different context, LLMs were used in combination with error minimization to describe cascade neural network architectures [3]).

An essential ingredient of the method is the differential $D\boldsymbol{f}$ of the continuous nonlinear map $\boldsymbol{f}$ linking the patterns. So far, only models with LLMs limited to rotations, i.e. $\det D\boldsymbol{f} = 1$, have been examined in the literature. Here we extend the model's scope to arbitrary deformations by making use of the most general form the differential can assume. In the following, we describe this generalization and discuss, as a first specific extension, the case of *conformal* local maps (local rotations and stretchings).

2 Local Linear Maps and Matching of Data Patterns

Following Aonishi and Kurata [2] we define deformations of data patterns as transformations produced by continuous nonlinear maps: Given an original pattern $I_1(\boldsymbol{x})$, $\boldsymbol{x} \in D_1 \subset \mathcal{R}^n$, it is assumed that there exists a function $\boldsymbol{f} : D_1 \mapsto D_2 \subset \mathcal{R}^n$ such that the deformed image $I_2(\boldsymbol{y})$, $\boldsymbol{y} \in D_2$, is obtained by $I_2(\boldsymbol{f}(\boldsymbol{x})) = I_1(\boldsymbol{x})$.

The key idea of local linear maps (LLMs) is to replace the global map $\boldsymbol{f}$ in (overlapping) neigbourhoods U_i of points $\boldsymbol{x}_i$ with its *linear* approximation, $f(\boldsymbol{x'} - \boldsymbol{x}_i) = (D\boldsymbol{f})|_{\boldsymbol{x}_i}(\boldsymbol{x'} - \boldsymbol{x}_i)$, and thus to describe the manifold defined by $\boldsymbol{f}$ by the piecewise continuous composition of its tangential hyperplanes at the points $\boldsymbol{x}_i$.

The comparison of the original and the image pattern involves checking the similarity of mutual feature vectors extracted by folding the patterns with suitable kernels $\psi(\boldsymbol{x})$, e.g. Gabor filters. Let $J_1(\boldsymbol{x}) = \int_{\boldsymbol{x}' \in D_1} I_1(\boldsymbol{x}')\psi(\boldsymbol{x} - \boldsymbol{x}')\, d^n\boldsymbol{x}'$ quantify a feature with the help of ψ in the original pattern in the neigbourhood of $\boldsymbol{x}$. It is easy to show, using the linear approximation for $\boldsymbol{f}$, that the feature associated with the neigbourhood of its image point $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$,

$$J_2(\boldsymbol{y}) = \int_{\boldsymbol{y}' \in D_2} I_2(\boldsymbol{y}')\psi(\boldsymbol{y} - \boldsymbol{y}')\, d^n\boldsymbol{y}' \;, \tag{1}$$

is related to the *original* data pattern through the modified kernel $\psi((D\boldsymbol{f})|_{\boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}'))\, \det((D\boldsymbol{f})|_{\boldsymbol{x}'})$, viz.

$$J_2(\boldsymbol{y}) = \int_{\boldsymbol{x}' \in D_1} I_1(\boldsymbol{x}')\, \psi((D\boldsymbol{f})|_{\boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}'))\, \det((D\boldsymbol{f})|_{\boldsymbol{x}'})\, d^n\boldsymbol{x}' \;. \tag{2}$$

In other words, the filter has to be transformed according to the local linear map $D\boldsymbol{f}$ taken at the original point $\boldsymbol{x}$ whose image is $\boldsymbol{y}$.

Given some pattern $I_2(\boldsymbol{y})$, we can compute feature vectors $J_2^k(\boldsymbol{y}_i)$ from (1) using filters ψ^k. The problem of deciding whether or not I_2 is the deformed image of an original pattern $I_1(\boldsymbol{x})$ then amounts to the problem of finding a map $\boldsymbol{f}$, with its local linear maps $D\boldsymbol{f}$, such that $J_2^k(\boldsymbol{y}_i)$ can equivalently be computed, for every k and i, from the *original* pattern I_1 using the filters transformed according to (2).

3 Theory

3.1 Diagonalisation of the Jacobian in 2-d

From now on we shall restrict ourselves to two-dimensional data patterns. In components and cartesian coordinates, $\boldsymbol{f} = (f_1(x, y), f_2(x, y))^{\mathrm{T}}$, the differential $D\boldsymbol{f}$ is represented by the Jacobian matrix

$$J = \begin{pmatrix} f_{1x}(x, y) & f_{1y}(x, y) \\ f_{2x}(x, y) & f_{2y}(x, y) \end{pmatrix} \;, \tag{3}$$

where, as usual, the index denotes the partial derivative with respect to the coordinate x or y. A novel aspect in our approach is that we make use of the fact that at every point the Jacobian in two dimensions can be brought into diagonal form,

$$R^{-1}(\gamma_2)\, J\, R(\gamma_1) \;=\; \Lambda \;=\; \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \tag{4}$$

with two appropriate rotations by angles γ_1 and γ_2, and two stretchings by factors λ_1 and λ_2. All four parameters are of course functions of position.

To prove this, consider orthonormal basis vectors $\hat{\boldsymbol{u}}_1, \hat{\boldsymbol{v}}_1$ defining a cartesian coordinate system at a given point $P(x, y)$, rotated by an angle γ_1 with respect to the (x, y) coordinate system. The Jacobian maps $\hat{\boldsymbol{u}}_1, \hat{\boldsymbol{v}}_1$ on two, in general, non-orthonormal vectors $\boldsymbol{u}_2, \boldsymbol{v}_2$ located at the image of P; however, for a special choice of γ_1 these two vectors can be made orthogonal, with the corresponding coordinate axes rotated with respect to the image (x, y) system by γ_2.

Then evidently $\hat{\boldsymbol{u}}_1 = R(\gamma_1)\,\hat{\boldsymbol{e}}_x$, and $\boldsymbol{u}_2 = \lambda_1 R(\gamma_2)\,\hat{\boldsymbol{e}}_x$, with $\lambda_1 = \|\boldsymbol{u}_2\|$. Since $\boldsymbol{u}_2 = J\hat{\boldsymbol{u}}_1$, we have $\lambda_1 R(\gamma_2)\,\hat{\boldsymbol{e}}_x = J\,\hat{\boldsymbol{u}}_1 = JR(\gamma_1)\,\hat{\boldsymbol{e}}_x$, hence $R^{-1}(\gamma_2)\,J\,R(\gamma_1)\,\hat{\boldsymbol{e}}_x = \lambda_1 \hat{\boldsymbol{e}}_x$. Similarly, $R^{-1}(\gamma_2)\,J\,R(\gamma_1)\,\hat{\boldsymbol{e}}_y = \lambda_2 \hat{\boldsymbol{e}}_y$, with $\lambda_2 = \|\boldsymbol{v}_2\|$. Thus $\hat{\boldsymbol{e}}_x$ and $\hat{\boldsymbol{e}}_y$ are eigenvectors of $R^{-1}(\gamma_2)\,J\,R(\gamma_1)$, with eigenvalues λ_1, λ_2, which proves (4). The values of γ_1, γ_2, λ_1, λ_2 can be determined in a straightforward way by setting up $R(\gamma_1)$, $R(\gamma_2)$ as 2-d rotation matrices and requiring the off-diagonal matrix elements of $R^{-1}(\gamma_2)\,J\,R(\gamma_1)$ to be zero.

3.2 Most General Form of Local Linear Maps

The most general form of a local linear map can now be gained by solving (4) for the Jacobian, $J \;=\; R(\gamma_2)\,\Lambda\,R(-\gamma_1)$, and decomposing Λ in the form

$$\Lambda = \sqrt{\lambda_1\lambda_2}\begin{pmatrix} \sqrt{\lambda_1/\lambda_2} & 0 \\ 0 & \sqrt{\lambda_2/\lambda_1} \end{pmatrix} = \bar{\lambda}\begin{pmatrix} \kappa & 0 \\ 0 & \kappa^{-1} \end{pmatrix}, \tag{5}$$

with $\bar{\lambda} = \sqrt{\lambda_1\lambda_2}$, $\kappa = \sqrt{\lambda_1/\lambda_2}$. We can then write for the Jacobian

$$J = \underbrace{R(\gamma_2)\begin{pmatrix} \kappa & 0 \\ 0 & \kappa^{-1} \end{pmatrix}}_{A} \cdot \underbrace{\bar{\lambda} R(-\gamma_1)}_{C}. \tag{6}$$

From this we see that in two dimensions J is the product of a *conformal*, i.e. *angle preserving*, map C (rotational stretching: each infinitesimal shape is uniformly blown up by $\bar{\lambda}$ and rotated by $-\gamma_1$, and an *area preserving* map A (rotational squashing: each infinitesimal shape is squashed by κ, with its area maintained, and rotated by γ_2). The decomposition (6) suggests in a natural way the following classification of local linear maps.

3.3 Generic Local Linear Maps

Four-Parameter Maps This comprises the most general case discussed above of two rotations and two anisotropic stretchings, or, alternatively, a conformal followed by an area preserving map.

Three-Parameter Maps This is the combination of one rotation (e.g. $\gamma_2 = 0$) plus anisotropic stretchings, which can also be considered as a conformal map followed by an area preserving squashing.

Two-Parameter Maps

a) Conformal Maps: $\kappa = 1,\ \bar{\lambda} \neq 1$ (i.e. $\lambda_1 = \lambda_2 = \bar{\lambda}$), $J = \bar{\lambda}\, R(\gamma_2 - \gamma_1) = \bar{\lambda}\, R(\theta)$ (angle preserving rotation and isotropic scaling). The components f_1 and f_2 must satisfy $f_{1x} = f_{2y},\ f_{1y} = -f_{2x}$ (Cauchy-Riemann equations). We note that every analytic function in the complex plane induces a mapping of this type.
b) Area-Preserving Maps: $\gamma_1 = 0$, $\bar{\lambda} = 1$ (i.e. $\lambda_2 = 1/\lambda_1$), and $\kappa = \lambda_1$ (rotation plus area preserving squashing). For f_1 and f_2 we have the constraints $f_{1x} f_{1y} + f_{2x} f_{2y} = 0$ and $\det J = f_{1x} f_{2y} - f_{1y} f_{2x} = 1$.
c) Anisotropic Stretchings: $\gamma_1 = \gamma_2 = 0$, and $\lambda_1 \neq \lambda_2$.

One-Parameter Maps

a) Rotations: These can be regarded as conformal mappings without stretchings, $\lambda_1 = \lambda_2 = 1$; at every point J has the form of a rotation matrix, that is f_1, f_2 fulfill the Cauchy-Riemann equations and $\det J = 1$.
b) Isotropic Stretchings: This is the simple case of no rotations, $\gamma_1 = \gamma_2 = 0$, and $\lambda_1 = \lambda_2$ at every point.

4 Method of Solution

Only the case of rotations has been discussed in the literature so far (Aonishi and Kurata [2]). An analysis of dynamic link matching in the context of local linear maps for all other generic cases is still lacking. However, the general procedure for determining the map $\boldsymbol{f}$ by matching local feature vectors $\boldsymbol{J}_1(x, y)$ in the original and $\boldsymbol{J}_2(x', y')$ in the deformed data pattern can be adopted from [2]:

1) Consider estimator functions [4] $\hat{\boldsymbol{f}}(x, y) = (\hat{f}_1(x, y), \hat{f}_2(x, y))^{\mathrm{T}}$.

2) Choose the type of generic local linear map which is to represent the differential $D\hat{\boldsymbol{f}}$; in the most general case it will depend on four parameters (6), $(D\hat{\boldsymbol{f}})(x, y) = A(\gamma_2(x, y), \kappa(x, y))\, C(\gamma_1(x, y), \bar{\lambda}(x, y))$.

3) To produce a topographic transformation between two data patterns I_1, I_2, formulate an appropriate cost function $C_{\mathrm{total}}[\hat{\boldsymbol{f}}, D\hat{\boldsymbol{f}}, \nabla\gamma_1, \nabla\gamma_2, \nabla\kappa, \nabla\bar{\lambda}]$.

4) Find the global minimum of the cost function by solving the time evolution equations $d\hat{f}_1(x, y)/dt$, $d\hat{f}_2(x, y)/dt$, $d\gamma_1(x, y)/dt$, $d\gamma_2(x, y)/dt$, $d\kappa(x, y)/dt$, $d\bar{\lambda}(x, y)/dt$ that result from a steepest gradient descent of the cost function.

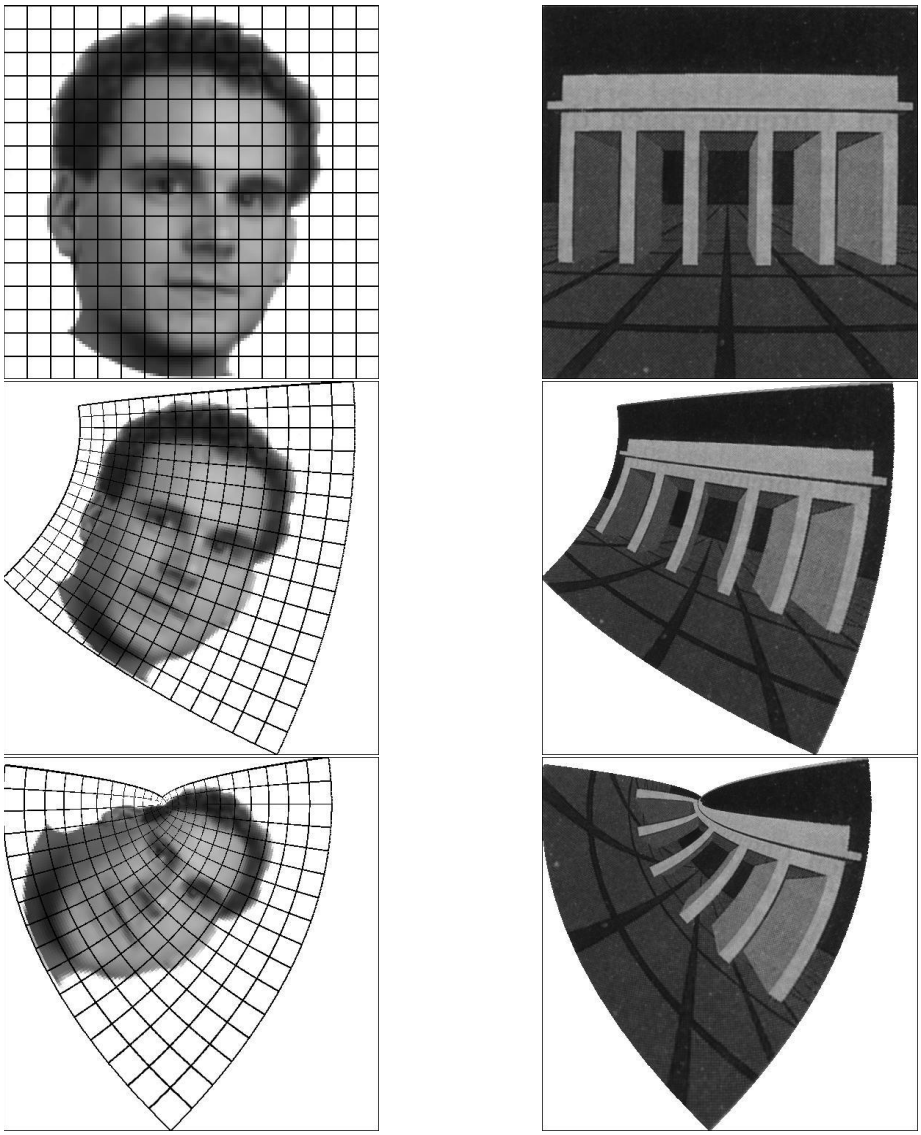


Fig. 1. Distortion of data patterns described by local linear maps. Example: conformal local linear maps (rotations and isotropic stretchings). The figure shows the effect of a conformal local linear map on an amorphous (left) and a well-structured (right) data pattern. Top: original data pattern ($\epsilon = 0$ in (7)), middle: $\epsilon = 0.0001$, bottom: $\epsilon = 0.005$. Note the smooth variation of the local rotation angles and stretching factors.

The choice of the cost function is crucial to finding the "best" global map $\hat{\boldsymbol{f}}$ and, with it, the "best" local linear maps. The cost function terms given in [2] can serve as a useful guide for the other generic cases. Obviously, the detailed discussion is beyond the scope of this presentation.

5 Example: Local Linear Maps as Conformal Maps

We will now consider the application of two-parameter maps, viz. conformal maps. To obtain the best local conformal map linking a given pair of data patterns, the cost function has to be minimized including the constraint that accounts for the Cauchy-Riemann equations. The implementation of this procedure is in progress, and detailed results will be published elsewhere. Here we will restrict ourselves to the illustration of the effects of conformal maps on the deformation of given data patterns. As an example, we show in Figure 1 the effects of the conformal mapping

$$\begin{pmatrix} f_1(x,y) \\ f_2(x,y) \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} + \epsilon \begin{pmatrix} x^2 - y^2 \\ 2xy \end{pmatrix} \tag{7}$$

on two data patterns. The smoothly varying local rotation angles and stretching factors can clearly be recognized from the deformation of the rectangular grid. In actual implementations, the best conformal local map must be determined from the original and the deformed data pattern using the procedure described in Section 4.

Acknowledgements. We thank G. Hornig for pointing out the diagonalization of the Jacobian.

References

1. Lades, M., Vorbrüggen, J. C., Buhmann, J., von der Malsburg, C., Würtz, R. P., Konen, W.: Distortion Invariant Object Recognition in the Dynamic Link Architecture. IEEE Trans. Comput. **42** (1993) 300–311
2. Aonishi, T., Kurata, K.: Extension of Dynamic Link Matching by Introducing Local Linear Maps. IEEE Transactions of Neural Networks **11** (2000) 817–822
3. Littman, E., Ritter, H.: Cascade LLM Networks. In: Aleksander, I., Taylor, J. (eds.): Artificial Neural Networks 2. Elsevier Science Publishers B. V., North Holland (1992) 253–257
4. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol. Cybern. **43** (1982) 59-69

Learning the Detection of Faces in Natural Images

Alexander Heinrichs[1], Christian Eckes[2], Rolf P. Würtz[1], and Christoph von der Malsburg[1,3]

[1] Institut für Neuroinformatik, Ruhr-Universität Bochum
Universitätsstr. 150, D-44780 Bochum, Germany
[2] Fraunhofer Institute for Media Communication,
Schloss Birlinghoven, D-53754 Sankt Augustin, Germany
[3] Lab for Computational and Biological Vision
University of Southern California, Los Angeles (CA), USA

Abstract. We present a two-stage face-finding system as a combination of labeled graph matching and statistical learning. The data format for both stages consists of vectors of the responses of Gabor wavelet filters. Graph matching is used to detect possible locations of faces that we call hypotheses. These typically contain many false positives. The graphs at the found locations are then reinterpreted as vectors, which can be used as input for different statistical learning methods. The methods used here are K-Nearest-Neighbour and the Support Vector Machine with the latter being more efficient.

1 Introduction

Face recognition systems based on elastic graph matching with Gabor wavelet preprocessing have shown outstanding competitive performance (see [PMRR00]) but are less suitable for detecting faces in complex scenes in the first place. Statistical investigations of patterns of faces and face-like objects are needed to overcome such limitations. Examples of successful systems following this approach include a trained multi layer perceptron performing face detection with feature vectors directly derived from grey-level images[RBK98]. Since Gabor Wavelet preprocessing and Support Vector Machines show outstanding performance in computer vision and statistical learning, it makes sense to combine both approaches in designing algorithms for face validation.

The system we propose here, is based on the system developed by Loos and Wieghardt [WL01], which we call face-finder in the following. It detects a chosen number of faces from any given image by using a simplified version of *bunch graph matching* (see [WFKvdM97] for details) and matching with a skin-color template. It delivers so called *hypotheses* of face-positions. The main idea of the new system is to combine the old version with an additional validation stage based on statistical learning to check the output of the face-finder and to improve the results.

The statistical learning methods used in this work are K-Nearest-Neighbour and linear and non-linear Support Vector Machines for classification. A detailed description of the hypothesis generation and the underlying bunch graph matching algorithm is available at [WL01]. The images used for this work are from varying sources (see [LJL98], [SH94], [TP91], [RBK98] and [BHK97] for details). Thanks to the colleagues for allowing us to use them for scientific purposes.

2 Classification of the Hypotheses

2.1 Preparation for the Classification

When the hypotheses are found, a graph is laid on every found region and the Gabor responses at the nodes are stored. The graph has the same structure as the bunch graph but just one jet stored at every node.

The samples used in statistical classifiers are vectors. Therefore, the graph structure of a hypothesis is flattened to a 640-dimensional vector. (There are 40 values in each jet and 16 nodes in one graph.) The vectors are normalized in three different ways to test the influence of the normalization on the classification. In the first dataset the vectors are normalized as a whole, in the second the single jets are first normalized and then written to a vector and in the third the vectors remain unnormalized. Normalization aims at making the representation invariant against contrast changes.

Now the hypotheses are labeled manually. The classes of faces, non-faces and unsure elements are established. The *unsure*-class contains faces which are poorly located by the face-finder due to, e.g., depth rotations not covered by the purely frontal bunch graph. Depending on the experiment, this unsure-class is treated as face or non-face or left out completely to study the influence of these hard-to-learn examples. Thus, the classification task remains a two-class-classification. In the last step of the preparation, the dataset is reordered randomly and every other value is part of the training- or test-set, respectively. Altogether a set of 3115 hypotheses was used, which was build by 855 faces, 392 unsure cases and 1868 non-faces.

2.2 Classification methods

The classification methods used in this work are K-Nearest-Neighbour, linear and non-linear Support Vector Machines for classification. As a relatively simple method, KNN does not offer optimal performance for classification, but can give an impression of the quality of the dataset and the possible performance of better classification-methods. In this work KNN has been applied with the Leave-One-Out-method, i.e., all data except one is used as training-set and the one vector is classified.

The concept of the Support Vector Machine was first introduced by Vapnik in [Vap95] and has been used in many applications with remarkable success. The method can be used for classification and regression of data but in this work, just

k	no Unsure			Unsure as Non-faces	Unsure as Faces
	Normalized Jets	Normalized Vectors	no Normalization		
1	12.67	14.22	15.70	12.45	13.55
3	12.52	13.07	15.83	11.60	12.74
5	11.86	12.04	15.47	11.53	12.41
7	11.90	12.26	14.80	11.16	12.12
9	11.75	11.72	14.96	11.64	12.38
11	12.05	12.17	15.12	11.75	12.56
21	12.63	11.88	16.08	12.93	14.03
31	13.04	11.65	16.78	14.36	15.24
51	13.37	11.69	17.56	15.39	16.45
101	14.73	12.10	18.39	16.82	18.69

Table 1. Error-rates of the tests with the KNN-Method.

classification is used. In [Sch97] a good introduction to the theory of Support Vector Machines is given. There are several algorithms and implementations of the SVM-method available. In this work the implementation SVM-Light by Joachims (see [Joa99] for details) is used.

3 Results

In this paper, only few of the computed results can be presented. For a detailed view on the results, please have a look at [Hei01] (in German).

3.1 Tests with KNN

Table 1 shows the error-rate for classification with varying values of k, the number of neighbours considered. The variation of the normalization type is shown without the unsure-cases being used. One can see that for small values of k the type of normalization makes just a small difference. With a bigger k the error-rate increases for all normalization types. This effect was the same for all three methods of treatment of the unsure class, but smallest without the unsure-class. Tests were done with even bigger values of k. But as the size of the dataset has the same magnitude as k, these results depend on the chosen dataset.

Table 1 also shows the error-rates for classification with the unsure-class treated in the three different ways. Notable is the difference between the error-rate *Unsure as faces* and the other error-rates. Looking closer on the results, one can see that especially the classification of the negative examples is worse in this case. A reason could be that for the correct classification of non-faces more data is required than for the classification of faces. Most of the faces are in a relatively small area of the feature space while the non-faces are scattered around everywhere.

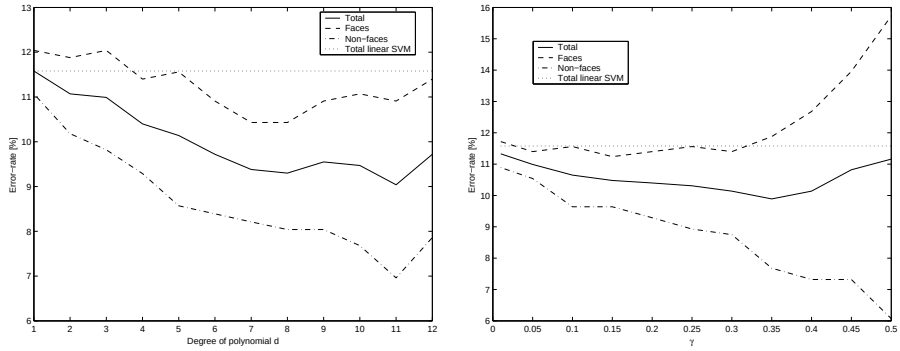


Fig. 1. Error rate vs. degree of polynomial (left) and factor γ (right)

3.2 Tests with SVM

The following results from the experiments are just a small fraction of all results and are chosen for several reasons. They all are created with the unsure elements counted as faces and normalized jets. The normalization of jets is chosen because the tests with KNN showed that this preprocessing is more stable than the other two versions. The unsure elements are counted to the faces-class, because this fits better to human perception (see figure 2 for examples). As mentioned before, the unsure elements tend to be faces, which are not found exactly. Human observers would have no problem in classifying them as faces and so the system should learn to classify them in this way, too.

In general the performance of SVM is better than KNN regardless of the chosen parameters and kernel function. On the left, figure 1 shows the error-rate for the polynomial SVM and the linear SVM. The exact value of the error-rate for the linear SVM is 11.6%. The use of the polynomial kernel is justified, because the error-rate can be decreased to 9.0%. The best value of the degree of the polynomial is 11, which shows, that a relatively complex separation plane divides the data better. Especially the error-rate of the non-faces decreases up to this value. When the degree gets still higher, the performance decreases because of over-fitting of the separating hyper-plane.

Figure 1 also shows the classification performance for the use of a Gaussian kernel on the right. The parameter γ, which is varied here, controls the complexity of the separating plane – a high value of γ leads to a more complex plane and more support vectors. With decreasing γ the performance gets better especially for the negative examples. The performance for the faces stays the same or gets even worse. The faces lie in a relatively small area of the space and can be classified properly with few support vectors. The scattered non-faces can be classified better with more support vectors. But if the number of non-face support vectors increases, the probability of a positive example to be misclassified rises as well.

The use of a polynomial function delivers the best performance for this problem. Additional tests have shown that the performance of the system with Gaus-

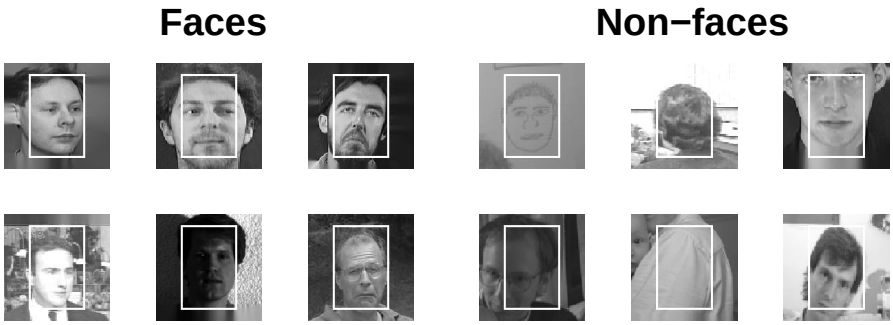


Fig. 2. Examples of hypotheses, which became support vectors with the linear SVM.

sian kernel is far more dependent on the normalization of the vectors. It is difficult to adjust the parameter γ to get a good performance for the unnormalized vectors, since these are far more scattered in the space.

4 Discussion

The classification of the hypotheses of the face-finder into faces and non-faces with the Support Vector Machine delivers satisfactory results. The delivered negative hypotheses are very similar to faces in the Gabor responses and so the shown performance is remarkably good.

Certainly the shown method can be developed further in different ways. On one hand, the performance of the hypothesis-finder might be increased by several steps. The concept of the bunch graphs might be developed further to make learning possible, while the system is working. At the current state a lot of human knowledge has to be put into the system to make it work and it is desirable to let the system start with little knowledge delivered by a human and learn independently how to increase the performance.

The performance of the SVM will increase with the amount of data available for training. Since manual labeling of these amounts of data (>10000 images) is awkward, ways have to be found for confident automatic or semi-automatic labeling. Like for the face-finder, it might be possible to start with knowledge provided by humans. This knowledge has to be increased automatically by the system to increase the performance.

As Schoelkopf has shown in [Sch97] it is possible to increase the performance on new data by retraining a SVM with just the previously misclassified data and the old support vectors. So it is possible to store only the misclassified data and the support vectors and still keep all necessary information.

For some additional insight into the form of the face class the support vectors after training have been inspected (see figure 2 for some examples). Notably many of the support vectors of the *non-face*-class still look pretty much like faces. This may indicate that the class boundaries are not very sharp or that the manual labeling has been rather conservative.

5 Acknowledgments

Funding by the German Federal Minister for Science and Education under the project LOKI (01 IN 504 E 9), the European Commission in the Research and Training Network MUHCI (HPRN-CT-2000-00111), and through the "Körber prize for European Science" awarded to C. von der Malsburg in 2000 is gratefully acknowledged.

References

[BHK97] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[Hei01] Alexander Heinrichs. Einsatz statistischer Lernmethoden zum Finden von Gesichtern in natürlichen Bildern. Internal Report IRINI 2001-07, Institut für Neuroinformatik, Ruhr-Universität Bochum, 2001.

[Joa99] T. Joachims. Making large-scale svm learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT-Press, 1999.

[LJL98] A. Loui, C. Judice, and S. Liu. An image database for benchmarking of automatic face detection and recognition algorithms. In *Proc. IEEE International Conference on Image Processing*, 1998.

[PMRR00] P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 22(10):1090–1104, 2000.

[RBK98] H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.

[Sch97] B. Schölkopf. *Support Vector Learning, Dissertation.* R. Oldenbourg Verlag, Munich, Germany, 1997.

[SH94] F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. In *Proc. 2nd IEEE Workshop on Applications of Computer Vision*, 1994.

[TP91] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.

[Vap95] V. Vapnik. *The nature of statistical learning theory.* Springer Verlag, 1995.

[WFKvdM97] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[WL01] Jan Wieghardt and Hartmut S. Loos. Finding faces in cluttered still images with few examples. In Georg Dorffner, Horst Bischof, and Kurt Hornik, editors, *Artificial Neural Networks – ICANN 2001*, volume 2130 of *Lecture Notes in Computer Science*, pages 1026–1033, Vienna, Austria, 2001. Springer Verlag.

Differential Processing of Facial Motion

Tamara L. Watson[1], Alan Johnston[1], Harold C.H Hill[2] and Nikolaus Troje[3]

[1]Department of Psychology, University College London, Gower Street, London WC1E 6BT
[2]ATR Human Information Science Laboratories, Kyoto, 619-0288
[3]Fakultät für Psychologie, Ruhr-Universität-Bochum, 44780 Bochum

Abstract. To investigate viewpoint dependence in dynamic faces an avatar was animated using actors' movements. In Experiment 1 subjects were shown a full-face animation. They were then asked to judge which of two rotated moving avatars matched the first. Test view, orientation and the type of motion were manipulated. In a second experiment subjects were shown two views of the same facial animation and were asked which of the two avatars was the same as the initial animation. Initial views could be rotated to 15° and 45° or 45° and 75° while test views were presented at 30° or 60°. Learnt view, test view, orientation and type of movement (rigid + non-rigid vs non-rigid) were manipulated. Both experiments and movement conditions produced an advantage for upright over inverted matching demonstrating subjects were encoding facial information. Non-rigid movement alone showed no effect of view for both experiments demonstrating viewpoint invariance. Rigid and non-rigid movement presented together produced a decline in performance for larger test rotations in Experiment 1, while Experiment 2 produced a differential advantage for 30° test rotation when initially viewed upright faces were rotated to 15° and 45° however no difference was found in the 45° and 75° condition or with inverted faces. These experiments suggest that non-rigid facial movement is represented in a viewpoint invariant manner whereas the addition of rigid head movements encourages a more viewpoint dependent encoding when the initial orientation of the head is not rotated further than the half profile (45°).

What role does motion play in the recognition of faces? Two types of motion, rigid transformations of the head and non-rigid deformations that occur during speech and changes in expression, are available to the viewer during social interaction. Research to date suggests that rigid motion of a head does provide beneficial information for the viewer. Pike et al. [1] have shown that this additional motion information presented at learning can enhance recognition. It is suggested that this advantage is affected by the ability to build up or access a 3-dimensional representation. The extra structural information provided by the rigid transformational motion of the head offers more opportunity to encode or access this information. However it is rare that when we are introduced to a person we see their face moving in the highly controlled way that was adopted by Pike et al. [1]. During most social interaction we will also be exposed to the face moving in a non-rigid manner.

The advantages of non-rigid motion for recognition have been the subject of debate. It has been shown that a degraded representation of a face will benefit from

the addition of non-rigid motion particularly for faces the viewer is familiar with. Knight and Johnston [2] have demonstrated that recognition of degraded famous faces will be significantly enhanced by the addition of non-rigid motion. Lander, Christie and Bruce [3] have demonstrated the same advantage with degraded famous faces. Christie and Bruce [4] have studied the effects of presenting non-rigid motion at training and at test with unfamiliar faces. They found no advantages for presenting motion at training or at test and suggest that non-rigid motion may only be beneficial when accessing existing representations.

Recently Thornton and Kourtzi [5] have used a sequential matching task rather than a recognition task in the study of non-rigid facial motion. They demonstrate that presentation of a short video sequence aided matching when the face differed in expression or viewpoint between prime and test images. The demonstration of an advantage in sequential matching of unfamiliar faces after presentation of a face moving non-rigidly in this study is interpreted with the view that mechanisms responsible for representing change over time are established and maintained in working memory and show little transference to long term memory over the course of the study.

All of the above studies have presented spatial layout cues alongside motion cues and have therefore not studied the role of facial motion alone. The question of whether facial motion can be represented independently of spatial cues remains open. However, Hill and Johnston [6] have shown that both rigid head movements and non-rigid head movements in the absence of spatial cues provide sufficient information to allow observers to categorize faces on the basis of both identity and gender. On the basis of differences between accuracy of categorization depending on the type of motion, Hill and Johnston [6] suggest that rigid movements are idiosyncratic and provide the basis for performance in identity categorization while non-rigid movements provide independent cues to speech and expression. These results would appear to complement the findings discussed above in that a more permanent representation is possibly mediated by encoding rigid motion while speech and expression are both encoded in a more transient manner.

The recognition of static faces has typically been found to be viewpoint dependent. Results of studies such as that by Hill, Schyns and Akamatsu [7] suggest that when a single view is presented during a learning stage, recognition of the same face from other views is impaired. They also found that the addition of cues that do not vary over view, such as facial colouring, greatly enhanced the accuracy of the results to the extent that learning presentation times need to be reduced. Recognition for the reduced presentation time was also found to be view dependent for conditions except in the case of the ¾ learnt views. These results suggest that generalized prior knowledge of the 3-dimensional structure of faces does not allow a view invariant representation of a face to be accessed when generalizing from a single static view.

As non-rigid facial motion is specifically a property of the object in motion it cannot be mimicked by movement of the viewer in the same way that rigid transformations can. Since this non-rigid motion is a change in the intrinsic shape of an object, it would make sense for the visual system to encode the motion in a viewpoint independent way if possible.

The first experiment was designed to assess view dependence when matching non-rigid facial movement as opposed to both rigid and non-rigid movement together from a full-face view.

Stimuli for both experiments reported consisted of a total of 64 animations based on motion capture recordings of 8 males and 8 females, each telling 4 question and answer type jokes. Recordings were made with an eight camera Oxford Metrics' Vicon motion capture system with the cameras placed in a semicircle at different heights in front of the head. Forty markers were used to capture facial movement and a headband with 4 markers was used to capture rigid movements. The resulting motion information was used to animate an average 3-dimensional facial model created from 100 male and 100 female faces [8]. Animation of the 3d model was achieved in Famous Animator where 'areas of influence' around each marker placed on the face inherit the movement of the marker (see also [6]). As no eye movements were captured the eyes were made to "look at" a point straight ahead of the face. The three-dimensional head model was texture mapped with a corresponding average texture and the resulting animated sequences rendered using 3DS Max. Two versions of each sequence were rendered; one with just non-rigid facial movements and the other with both types of movements combined.

Two groups of 20 subjects were presented with animations containing only non-rigid motion or rigid and non-rigid together. During one trial participants were first shown a learning animation sequence oriented at 0° (where 0° is a full face and 90° a profile). This was followed by a target and distracter animation presented sequentially at an orientation in depth of 0°, 15°, 30°, 45°, 60°, 75° or 90°. Participants were asked to indicate which animation was shown in the learning stage. The target animation was the same sequence as the learning animation while the distracter was randomly chosen with the constraint that it contains an actor telling the same joke as the target stimulus. Both target and distracter animations were shortened such that the video sequence would start at a random point within the first half of the animation and run for half the length of the full animation. Shortening the animation was required in order to lower performance from ceiling. Each animation could only be viewed once and all animations were required to have been viewed before response. Subjects controlled the speed of presentation.

Faces were also presented upside down as a control in order to assess the likelihood that subjects were utilizing extraneous cues in order to carry out the task. It has been shown previously that presenting inverted facial motion reduces the accuracy of gender and identity judgments suggesting that upright facial motion is represented in a object-motion encoding system specialized for faces [6]. Inversion constituted a within-subjects condition. During the upright condition all faces were presented upright. During the inverted condition all faces were presented upside down. Initial orientation was randomized. Each condition contained 64 trials.

A within subjects analysis was carried out for each type of motion. Data are shown in Table 1. The effect of test viewpoint for non-rigid motion was not found to be statistically significant, $f(6,114)=2.163$, $p>0.05$ although a significant overall effect of inversion was found, $f(1,19)=5.834$, $p=0.03$. Rigid and non-rigid motion together displayed a significant effect of test rotation $f(6,114)=2.311$, $p=0.04$ which was found to produce a linear trend, $f(1,19)= 14.468$, $p<0.01$. An effect of inversion was also found, $f(1,19)=5.819$, $p=0.03$.

Table 1. Mean percent correct for Experiment 1

	Non-Rigid Motion						
	0°	15°	30°	45°	60°	75°	90°
Upright	82.5	77.2	77.8	75.0	78.3	74.5	73.9
Inverted	76.5	65.0	78.3	67.8	68.1	71.1	77.2
	Rigid + Non-Rigid Motion						
	0°	15°	30°	45°	60°	75°	90°
Upright	82.8	82.1	81.5	77.2	74.1	75.3	69.8
Inverted	76.1	73.5	74.1	73.5	72.2	69.8	69.1

These results suggest that non-rigid facial motion is less viewpoint dependent than non-rigid and rigid facial motion presented together when generalizing from a full face view. Both the statistics and inspection of the data suggests that for upright non-rigid animation there is some decline in performance as test viewpoint rotates away from the target view. However this is not as pronounced as when all motion information is presented within the animation. The advantage displayed for upright animations is suggestive of a specialized face processing module that is not available when inverted animations are presented.

The next experiment was designed to incorporate a larger range of initial views. Non-rigid movement alone was presented to one group of 40 participants while another group viewed animations containing both rigid and non-rigid movements. Two different groups of subjects were shown 'learning faces' consisting of two different views of the same animation rotated by either 15° and 45° or 45° and 75° and were then tested on faces rotated to 30° or 60°. The target face was the same as the learnt face while the distracter was chosen randomly with the constraint that it would be at the same rotation as the test and of the same gender. Subjects were asked to indicate which animation had been shown in the learning stage. Both target and distracter animations were shortened and viewing conditions were as in Experiment 1. Inversion acted as an added within-subjects condition.

Table 2. Mean percent correct for Experiment 2

		Non-Rigid Motion		Rigid + Non-Rigid Motion	
Target Rotation		15° + 45°	45° + 75°	15° + 45°	45° + 75°
Upright Test	30°	81.3	81.6	85.0	88.1
	60°	79.4	81.2	76.3	90.6
Inverted Test	30°	74.3	75.0	74.4	78.1
	60°	70.0	77.8	71.9	78.4

Data are shown in Table 2. Non-rigid motion did not display an interaction between the initial and test rotation of faces $f(1,38)=0.152$, $p=0.70$, however an effect of inversion was found $f(1,38)=14.324$, $p<0.01$. The condition showing all available motion information displayed a significant interaction between the initial and test rotation of faces, $f(1,38)=9.215$, $p<0.01$. This interaction was not present when faces

were inverted, $f(1,38)=0.599$, $p=0.45$. Again an overall inversion effect was found, $f(1,38)=33.485$, $p<0.01$.

The results of Experiment 2 suggest that non-rigid motion when presented alone displays less viewpoint dependence than when all motion information is presented in an animation. Further investigation of the data also suggests that viewpoint dependence is stronger for the learnt animations closest to the full face view when compared to those closest to the profile in the full motion condition. This effect could be due to a larger angular difference between the more frontal views presented in this condition compared to the views closer to profile. The advantage for upright animations over inverted, as in Experiment 1, is suggestive of specialized processing of the upright facial motion.

The results reported here suggest that non-rigid transformations of a face may initially be encoded in a less viewpoint dependent manner than faces that are transforming and translating rigidly. It is also noted that while viewpoint dependence has been found in the latter case it does not seem as pronounced as that found when investigating view dependence in static faces. This may suggest that motion does have a large role to play in the perception of faces across different views. That non-rigid motion alone displays no view dependence unless teamed with rigid motion may suggest that the addition of rigid motion into the stimulus adds a layer of information that may make the representation of the motion that is formed less than optimal for the task at hand. Whether this is specific to the rigid motion or occurs as a combination of non-rigid and rigid motion is not clear from these experiments. The inversion effect displayed suggests that the results found in this study are due to specialized processing of the upright facial motion.

References

1. Pike, G., Kemp, R., Towell, N., Phillips, K.: Recognizing Moving Faces: The Relative Contribution of Motion and Perspective View Information. Visual Cognition 4. (1997) 409-437.
2. Knight, B., Johnston, A.: The Role of Movement in Face Recognition. Visual Cognition 4. (1997) 265-273.
3. Lander, K., Christie, F., Bruce, V.: The Role of Movement in the Recognition of Famous Faces. Memory and Cognition. 27. (1999) 974-985.
4. Christie, F., Bruce, V.: The Role of Dynamic Information in the Recognition of Unfamiliar Faces. Memory and Cognition 26. (1998) 780-790.
5. Thornton, I.M., Kourtzi, Z.: A Matching Advantage for Dynamic Human Faces. Perception 31. (2002) 113-132.
6. Hill, H.C.H, Johnston, A.: Categorizing Sex and Identity From the Biological Motion of Faces. Current Biology 11. (2001) 880-885.
7. Hill, H.C.H., Schyns, P.G., Akamatsu, S.: Information and Viewpoint Dependence in Face Recognition. Cognition 62. (1997) 201-222.
8. Vetter, T., Troje, N.: Separation of Texture and Shape in Images of Faces for Image Coding and Synthesis. Journal of the Optical Society of America 14. (1997) 2152-2161

A Neural Mechanism for Viewing-Distance-Invariance

Rüdiger Kupper and Reinhard Eckhorn

Philipps-University Marburg, Neurophysics Group, D-35037 Marburg, Germany

Abstract. We present a neural network mechanism allowing for distance-invariant recognition of visual objects. The term *distance-invariance* refers to the toleration of changes in retinal image size that are due to varying view distances, as opposed to varying real-world object size. We propose a biologically plausible network setup, based on the recently demonstrated spike-rate modulations by viewing distance, affecting large numbers of neurons in striate and extra-striate visual cortex. In this context, we introduce the concept of *distance complex cells*. We successfully implement the model in a computer simulation and investigate its response to changing view distances.

1 Introduction

Invariant object recognition means the ability of the human visual system to recognize familiar objects appearing in varying poses in the visual field, such as varying position, size, or three-dimensional view. It can be argued that positional invariance may mostly be achieved by fixational eye movements. Nonetheless, some sort of neural computation must be performed along the ventral pathway, to achieve invariance to size, view, or other transformations involving a change in retinal projection. Among these, transformation of size plays a special role, as it is characterized by changes in extent, but not in shape.

1.1 Size-Invariance vs. Distance-Invariance

Size-invariant object recognition demands closer investigation, regarding the possible causes that make a familiar shape appear in different sizes on the retina.

Viewing Distance. One reason for visual objects having varying retinal image size is, that the same or identical objects appear at different viewing distances. A possible source for this type of size variation is self-motion. The resulting images are perceived as being instances of the very same object even if there are huge differences in the extent of their retinal projections. We will refer to this type of invariant recognition as distance-invariance. It is unconsciously perceived.

Real-world Object Size. Another reason for varying retinal image extent can be that the observer is facing different objects of the same shape, but of different physical size, measured in real-world coordinates (e.g. a car and its toy model). These are perceived as being different objects, though possibly belonging to a common class.

Under normal viewing conditions, the two types of size variation can perceptually be well distinguished. The retinal image of a nearby toy car can match that of a far away real car. Nonetheless, perceptually the two objects are not confused. That is, *invariant recognition is achieved in the case of varying viewing distance, but not for varying real-world size.* We regard this as being of major importance. It means that physical size is an inherent object property used to distinguish between objects, and that it is derived considering the current viewing distance. To our knowledge, this is not accounted for by other models of size invariant object recognition, making use of neurally implemented two-dimensional image transformations [4], or of cascaded local feature pooling [3, 6]. There is, however, evidence, that the ability to distinguish these two types of size variation is also based on neural properties recently found in V1, V2 and V4 of monkeys [1, 2, 7].

1.2 Distance Estimation by the Visual System

Distance dependent modulation of single cell spike rates has been found to high abundance (64–85% of neurons) in visual cortical areas V1, V2, and V4, making it a very common property of cells at the lower levels of the ventral visual pathway [2, 7]. While cell properties like receptive field size and preferred two-dimensional stimulus properties (edge orientation, contrast, spatial frequency, etc.) stay unchanged, the cells exhibit a modulation of firing rate with fixation distance [1, 2]. The results can be interpreted as viewing distance being a further property coded by the respective neurons, in addition to their classical receptive field properties.

What functional purpose could the modulation of such large portions of neurons along the ventral pathway serve? We suggest, that viewing distance information is used to select, by sensitivity modulation, subsets of local feature detectors, which represent visual elements at a preferred viewing distance. The representation of a fixated object then is primarily made up from the responses of cells sensitive to the actual fixation distance.

2 Model

2.1 Extending the Concept of Hierarchical Visual Coding

Hierarchical models for object recognition adopt the view that increasingly complex features constitute the representation of objects [3, 6]. Our present model extends this concept by introducing the experimental finding of spike rate modulation by viewing distance into the hierarchy. Our model (Fig. 1) consists of, A) a linear neural chain representing the current fixation distance by a single activated blob, B) distance modulated feature detectors, C) distance complex cells, and D) an object knowledge base.

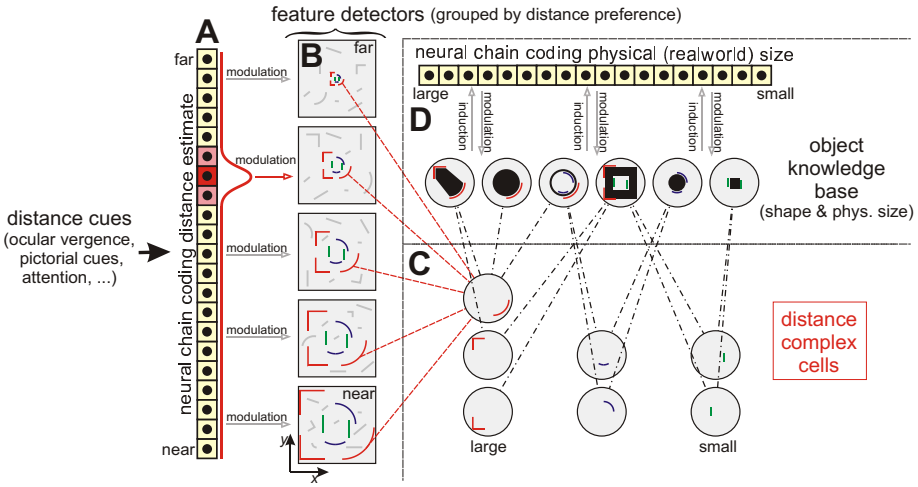


Fig. 1. Model architecture. Higher-order features are omitted for clarity. **A:** Neural chain coding distance by activation blob. **B:** Sets of feature detectors assigned to different viewing distances. **C:** Distance complex cells. **D:** Object knowledge base for shape and size, and neural chain coding the perceived physical size

A: Neural Chain Representing Distance. The exact origin and type of the distance signal is unknown. It can be provided from a variety of sources, including ocular vergence, lens accommodation, angle below horizon, or pictorial cues such as contrast, texture gradient and motion parallax. We model its action by a linear chain of coupled neurons, like a discretized one-dimensional neural field, in which the position of a single activation blob represents the current distance estimate of the ocularly fixated object (Fig. 1, A).

B: Modulation of Feature Detectors by Distance Signal. The retinal image is represented by the activation of low- and higher-level visual filters, each coding for their preferred features. Coding for distance is achieved by modulating their sensitivity by a distance signal [1, 2, 7] (Fig. 1, A and B). The distance tuning corresponds to the activation blob in (A:).[1]

C: Distance Complex Cells. Feature detector signals converge systematically onto next stage cells, yielding what we term *distance complex cells*. Their receptive field properties reflect the distance-variant transformation that a distinct visual feature undergoes, as the distance between observer and fixated object changes (Fig. 1, B and C, connections shown for one cell only). Throughout such a movement, the same distance complex cell would be kept activated.

[1] As experimental data [1, 2, 7] does currently not allow for exact shape estimation, we assume Gaussian tuning profiles.

D: Object Knowledge Base. The representation of objects as visual entities is based on the outputs of distance complex cells. Features are combined to yield representations defining not only the shape, but also the physical size of visual objects (Fig. 1, D). We use a second one-dimensional chain of neurons with a single activation blob to represent the physical size of a currently fixated object.

2.2 Operation of Model

Under real world viewing conditions the rich environment provides a distinct visual input, generally sufficient for a reliable distance estimate. As illustrated in Fig. 1, a subset of all feature detectors receives modulatory input from the neural chain representing the current viewing distance. Owing to this modulation, feature detectors of *appropriate* distance preference are facilitated and will predominantly represent the visual scene, while activity of *non-appropriate* detectors is diminished.[2] These detectors will feed the attached distance complex cells. The pattern of activated distance complex cells then activates a representation of correct shape and size in the object database.

Finally, activity in the different model modules carries information on identity (*shape* and *real-world size*), as well as the *distance* of the observed object. This mediates a stable percept of the three-dimensional scene, as the observer explores the environment in a series of saccades.

3 Results

We examined the model's response to hypothetical viewing situations, using a 3d-rendering system to compute retinal projections of an artificial three dimensional scene. Projection parameters such as position of lens and size of viewfield were set to match those of the human eye. Network input consisted of the retinal image, plus the current fixation distance (Fig. 2). No further information entered the network. Output was the activation of a topologically arranged set of distance complex cells.

3.1 Simulation Results

In a computer implementation, the model proves to generate size-invariant output from distance-varying views of an object. (Fig. 2). Although different viewing distances cause huge variations in retinal image size, the output of distance complex cells is largely independent of fixation distance. Figure 2 shows results for viewing distances of 30 and 60 m, but the network operates reliably over the full simulated distance range of 15–135 m. Note that the output is labelled in real-world coordinates: Objects of *different physical size will generate different output*, while representations *will not change with varying view distance*. The network thus reproduces the two viewing modes described in Sec. 1.1.

[2] Modulation affects the whole visual field, preserving spatial feature relations, but also causing false size transformations of peripheral objects not at the fixation distance.

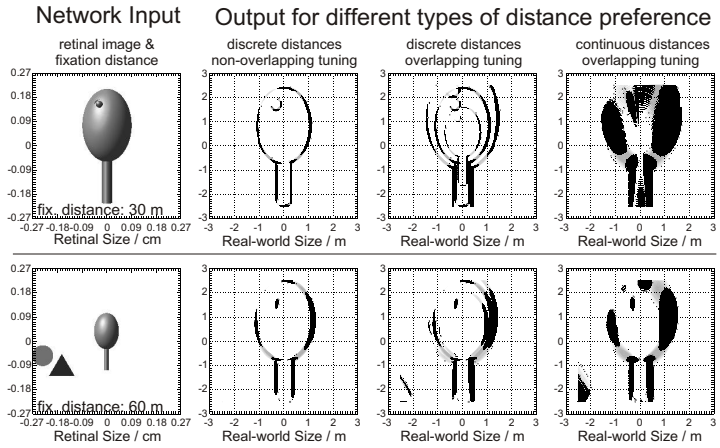


Fig. 2. Simulation results. Size-invariant output is generated from distance-varying stimuli. Note that output is labeled in real-world coordinates

Output quality depends on the spacing and width of the distance tunings: Non-overlapping tunings generate unique contours (Fig. 2, 2nd column), but their level of invariance depends on the density of distance sampling (i.e., the number of distance layers in Fig. 1, B). For broad, overlapping tunings, the level of invariance is high, but the network generates multiple ghost images of fixated objects (Fig. 2, 3rd and 4th column). Furthermore, the use of a single retinal frequency channel in the implementation causes increasing blur with fixation distance. The interdependence of retinal frequency tuning and distance modulation is currently subject to investigations in our laboratory [1], and will be incorporated into forthcoming versions of our model.

4 Discussion

The presented model belongs to the class of hierarchical models for object recognition. These are known to produce invariance when constructed accordingly [3, 6], but do so implicitly, losing information, e.g. on object size, position, and spatial relations among local features. Other models use control signals to set up image transformations [4], but act in the two-dimensional domain, unable to exploit the distance signal to gain information on physical object size. Our model can be seen as an extension to both strategies, using pooling operations guided by an explicit distance signal.

Based on the recent findings of firing rate modulation by fixation distance [1, 2, 7], we propose the existence of a new class of cells, exhibiting *complex* properties in the sense of being insensitive to feature transformations caused by change in viewing distance. Cells with receptive field properties that in several respects are similar to those of the hypothetical *distance complex cells* (Fig. 3)

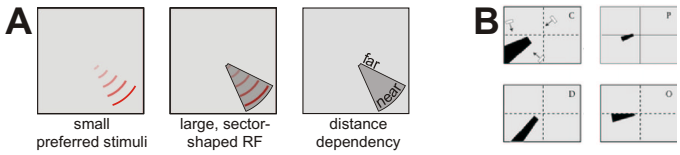


Fig. 3. RF-properties. **A:** Distance complex cells, as expected from model setup (Fig. 1). **B:** Cells in V4A are reported to have large, radially oriented, "comet-shaped" RFs. They prefer small stimuli and often respond to stimulation at appropriate distance only (Fig. taken from [5])

have recently been reported for the newly identified area V4A of monkey visual cortex [5]. These findings strongly encourage the further development of our model.

A possible drawback of our approach is the large number of required feature detectors. Detectors need to be present, which share the same preferred two-dimensional feature, but are sensitized for different viewing distances. The quality of invariance generation depends on the width and overlap of tuning profiles, as well as on the number of sampling points in distance. We will investigate, to what multiplicity detectors are required to allow for stable operation, and what constraints are imposed thereon by biological cell numbers. A radial gradient in spatial frequency preference could compensate for distance dependent blur [1].

Many more setups of our model can be investigated, including attention to distance and real-world size, attention to known objects, and operation in reduced cue environments (i.e., size changes with no distance signal available).

References

1. Brinksmeyer, H. J., Michler, F., Gail, A., Eckhorn, R.: Stimulus sensitivity in monkey visual cortex is modulated by viewing distance while spatial frequency tuning and receptive field size are not. Proc. Artificial Intel. **10**: Dynamische Perzeption (2002) (this volume)
2. Dobbins, A. C., Jeo, R. M., Fiser, J., Allman, J. M.: Distance modulation of neural activity in the visual cortex. Science **281** (1998) 552–555
3. Mel, B. W., Fiser, J.: Minimizing binding errors using learned conjunctive features. Neur. Comp. **12** (2000) 247–278
4. Olshausen, B. A., Anderson, C. H., van Essen, D. C.: A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J. Neurosci. **13(11)** (1993) 4700–4719
5. Pigarev, I. N., Nothdurft, H.-C., Kastner, S.: Neurons with radial RF in monkey area V4A: evidence of a subdivision of prelunate gyrus based on neural response properties. Exp. Brain Res. **145** (2002) 199-206
6. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature Neurosci. **2(11)** (1999) 1019–1025
7. Rosenbluth, D., Allman, J. M.: The effect of gaze angle and fixation distance on the responses of neurons in V1, V2, and V4. Neuron **33** (2002) 143–149

An iterative Bayesian technique for Dense Image Point Matching

Christian B. U. Perwass[1] and Gerald Sommer[2]

[1] Institut für Informatik, CAU Kiel, 24105 Kiel, Germany
christian@perwass.de, www.perwass.de
[2] Institut für Informatik, CAU Kiel, 24105 Kiel, Germany
gs@ks.informatik.uni-kiel.de

Abstract

We present a conceptually simple algorithm for dense image point matching between two multi-modal (e.g. color) images. The algorithm is based on the assumption that correct image point matches satisfy locally a particular statistical distribution. Through an iterative evaluation of a local probability measure, global constraints are taken into account and the most likely set of image point matches is found. An advantage of this approach is that no information about the camera geometries, as for example the epipoles, has to be known. Therefore, the algorithm can be used for stereo matching and optic flow.

1 Introduction

The basic idea behind all optic flow and stereo matching algorithms is, that if two images are projections of the same 3D-scene taken from slightly different positions or at slightly different times, then certain properties of corresponding pixels are invariant. However, it is not necessarily the case that a pixel in one image can be identified with exactly one pixel in the other, since rigid objects may appear shrunk or grown in different projections. Furthermore, parts of a 3D-scene that can be seen in one projection may be occluded in the other. The transformation between two images related by optic flow or stereo, is therefore more like a homotopy, as Florack et al. [1] point out, than a vector field. Nevertheless, a vector field is what we need in most applications. Therefore, in general an assumption is made about the invariant properties of corresponding pixel, which approximates nearly invariant properties of the underlying homotopy.

The invariant properties which are typically identified are those of pixel color and pixel neighborhood structure. Algorithms differ in how they model these invariances and the method employed in identifying corresponding pixels using the assumed invariant properties.

Some different types of approaches are for example: feature based methods (e.g. [2]), pixel labelling methods (e.g. [3, 4, 5]) and Bayesian methods (e.g. [6, 7, 8, 9]).

Bayesian methods have the advantage of clearly stating the invariance assumptions made about corresponding pixels by defining priors on the parameters of the system. Markov random field (MRF) approaches as described in [10] play an important role in this context [11]. The details of the different Bayesian approaches to dense image point matching are quite varied. However, typically they do not assign a single disparity label to a pixel but a discrete probability distribution function (pdf) over a set of disparities. Although this might, at first, seem to violate the often used uniqueness assumption as stated by Marr and Poggio [12], one can always define the final disparity to be the expectation value of the pdf. The advantage of defining a discrete pdf is that, in effect, we can test a number of hypotheses concurrently and eventually extract the most likely one. Finding the set of disparities which maximizes an appropriately defined probability measure then gives the answer to the correspondence problem. Such a maximization may be done iteratively or through a global maximization scheme.

In this paper we also follow a Bayesian approach which is based on an idea we published previously [13] using different mathematical tools. A detailed discussion of our approach, including a number of experiments, can be found in [14]. Our approach is similar to [15] but differs in the implementation of the pixel invariance properties. Where they use a MRF approach to enforce a smooth disparity space, we follow the idea that the distribution of correct pixel matches can locally be described by a particular pdf, whereas wrong match candidates are uniformly distributed. Through an iterative evaluation of a local probability measure, local matching constraints are propagated through the image, such that global constraints are taken into account. Although, occlusion is not modelled explicitly, half-occluded pixels are either given two different disparities simultaneously, or they are matched onto the nearest matchable pixel. That is, the algorithm does not break down in the presence of occlusion.

2 Theory

In the model we develop, we are not interested in the exact camera geometry. We simply assume that we are given two images A and B whose pixel are correlated in as far as they represent the same scene, albeit from a different point of view (stereo matching) or at a different time (optical flow). The only constraints we can invoke then are pixel similarity and an ordering constraint.

We assume that correct image point matches satisfy a particular statistical distribution whereas incorrect matches are equivalent to noise and are uniformly distributed. We are looking for an iterative procedure that amplifies those pixel that satisfy the appropriate distribution and subdues the others. We can only give a short overview of the algorithm's derivation here. For a detailed account see [14].

First of all we need a measure for pixel similarity. This measure has to express the likelihood that two pixels were created by the same element in a scene, without taking into account any neighboring pixels. Such a measure therefore

will be based on a pixel's color, but may also include any other local property like the local scale or local phase. We will denote this measure by $s(\mathbf{x}_A, \mathbf{x}_B)$, where $\mathbf{x}_A$ denotes a pixel position in image A and $\mathbf{x}_B$ a pixel position in image B.

Using $s(\mathbf{x}_A, \mathbf{x}_B)$, we can evaluate for each pixel in image A its similarity to the pixels within an area of image B where we expect the correct match to lie. We will also call this a test patch. That is, each pixel in image A has associated with it a probability distribution giving its matching likelihood to a set of pixels in image B. Our goal is to minimize the entropy of these probability distributions, i.e. to minimize the match uncertainty.

In order to do this, the pixel similarity measure alone is not enough. We also have to take into account a structural constraint. We do this by assuming that the local distribution of pixel matches takes on a particular form. This becomes the prior distribution in our derivation, denoted by $h(\mathbf{x}_A, \mathbf{x}_B, \mathbf{y}_A, \mathbf{y}_B)$. That is, given an assumed pixel match $(\mathbf{x}_A, \mathbf{x}_B)$ and a particular neighbor $\mathbf{y}_A$ of $\mathbf{x}_A$, $h(\mathbf{x}_A, \mathbf{x}_B, \mathbf{y}_A, \mathbf{y}_B)$ gives the a priori probability distribution for $\mathbf{y}_B$ being a correct match of $\mathbf{y}_A$.

It can be shown that the probability of $(\mathbf{x}_A, \mathbf{x}_B)$ and $(\mathbf{y}_A, \mathbf{y}_B)$ being two neighboring pixel matches is then given by

$$\begin{aligned} &P(\mathbf{X}_B = \mathbf{x}_B, \mathbf{Y}_B = \mathbf{y}_B | A, B, \mathbf{X}_A = \mathbf{x}_A, \mathbf{Y}_a = \mathbf{y}_A) \\ &\quad = s(\mathbf{x}_A, \mathbf{x}_B) s(\mathbf{y}_A, \mathbf{y}_B) h(\mathbf{x}_A, \mathbf{x}_B, \mathbf{y}_A, \mathbf{y}_B). \end{aligned} \tag{1}$$

The probability measure on which we base our match decision is the following. Assuming $(\mathbf{x}_A, \mathbf{x}_B)$ are a correct match, then for a given neighbor $\mathbf{y}_A$ of $\mathbf{x}_A$ we say that the most likely match $\mathbf{y}_B$ of $\mathbf{y}_A$ is the one where the data best satisfies the prior distribution of neighboring matches. That is we are looking for the estimator $\hat{\mathbf{y}}_B$ given by

$$\hat{\mathbf{y}}_B = \arg\max_{\mathbf{y}_B} \left(\frac{P(\mathbf{X}_B, \mathbf{Y}_B = \mathbf{y}_B | A, B, \mathbf{X}_A, \mathbf{Y}_A)}{\max_{\mathbf{y}} P(\mathbf{X}_B, \mathbf{Y}_B = \mathbf{y} | \mathbf{X}_A, \mathbf{Y}_A)} \right). \tag{2}$$

The effect of this is that if for a particular set $(\mathbf{x}_A, \mathbf{x}_B, \mathbf{y}_A)$ the corresponding $\hat{\mathbf{y}}_B$ maximizes the prior, then

$$P(\mathbf{X}_B = \mathbf{x}_B, \mathbf{Y}_B = \hat{\mathbf{y}}_B | A, B, \mathbf{X}_A = \mathbf{x}_A, \mathbf{Y}_A = \mathbf{y}_A) = s(\mathbf{x}_A, \mathbf{x}_B) s(\mathbf{y}_A, \mathbf{y}_B). \tag{3}$$

That is, the match probability depends solely on the pixel similarities.

What we really need to estimate is the probability of $(\mathbf{x}_A, \mathbf{x}_B)$ being a correct match. However, for each neighbor $\mathbf{y}_A$ of $\mathbf{x}_A$ we obtain a match probability estimate from $P(\mathbf{X}_B, \mathbf{Y}_B = \hat{\mathbf{y}}_B | A, B, \mathbf{X}_A, \mathbf{Y}_A)$. We therefore take the final match probability estimate of a pixel pair $(\mathbf{x}_A, \mathbf{x}_B)$ to be the expectation value of the set of probability estimates for all eight neighbors of $\mathbf{x}_A$.

$$\begin{aligned} &P(\mathbf{X}_B = \mathbf{x}_B \,|\, A, B, \mathbf{X}_A = \mathbf{x}_A) \\ &\quad = \rho \; s(\mathbf{x}_A, \mathbf{x}_B) \; \tfrac{1}{8} \textstyle\sum_{\mathbf{y}_A} \max_{\mathbf{y}_B} \; s(\mathbf{y}_A, \mathbf{y}_B) \; \dfrac{h(\mathbf{x}_A, \mathbf{x}_B, \mathbf{y}_A, \mathbf{y}_B)}{\max_{\mathbf{y}} h(\mathbf{x}_A, \mathbf{x}_B, \mathbf{y}_A, \mathbf{y})}, \end{aligned} \tag{4}$$

where ρ is a normalization constant and the sum over $\mathbf{y}_A$ goes over all eight neighbors of $\mathbf{x}_A$.

Evaluating the probability measure from equation (4) only once, will not give us a final match result. In order to minimize the entropy of the match probability distributions, we have to apply this measure iteratively. This distributes local match information throughout the image. It also means that homogeneous areas are matched according to the match constraints contained in their surroundings. In [14] we have shown that such an iteration converges. Note that this iterative procedure can be regarded as a recurrent neural network, whose equilibrium state gives the match result.

Half-occluded pixels, i.e. pixels that appear in one image but not in the other, have not been treated explicitly. However, by using a bidirectional matching scheme, the matching process is stabilized in the presence of half-occluded pixels.

3 Experimental Results and Conclusions

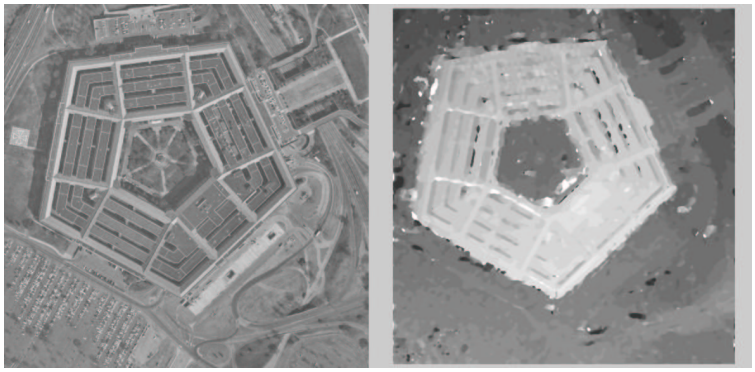


Fig. 1. Left image Pentagon example with evaluated disparity map.

In order to run the algorithm we have to set five parameters: the number of iterations to perform, the test patch size, the mean pixel displacement, the standard deviation of the ordering constraint σ_h and the standard deviation of the pixel similarity function σ_s. The mean displacement is basically an approximate pixel match. This is easy to find for optical flow, since we assume that corresponding pixel are almost at the same position. For stereo correspondence this initial match will have to be set by some other means. Finding the best number of iterations could be automated by stopping the algorithm once it has converged. The test patch size has to be set such that the correct match is always included. Here we have to make an assumption about how much we expect the pixels to have moved. The parameters σ_h and σ_s only change details of the final match result. They do not have to be changed for different images.

The Pentagon stereo pair was provided by CMU/VASC. Here we matched an area of 500×500 pixels with a test area size of 21×1. Figure 1 shows the first of the two Pentagon images together with the evaluated disparity map after 20 iterations. It can be seen that the algorithm works quite well for stereo matching on rectified images.

We used the Yosemite sequence created by Lynn Quann at SRI to test the algorithm in an optic flow setting. We matched the lower part of the first two images of the sequence, since no ground truth is available for the cloud region. The image dimensions were 315×177 pixels, the test patch size was 7×7 pixels. The parameters σ_s and σ_h had the same values as in the Pentagon example. We performed 20 iterations which took approximately 150 seconds on an AMD Athlon XP 1800+ (1.53 GHz) running Windows XP. The algorithm runs about twice as fast if we do not perform bidirectional matching, which stabilizes the algorithm in the presence of occlusion. Note that the implementation of the algorithm was experimental and not optimized for speed.

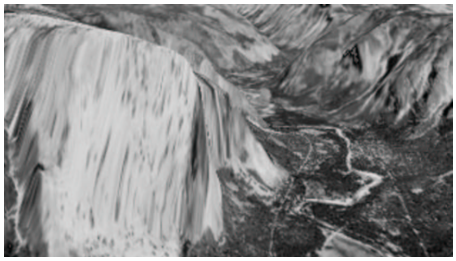


Fig. 2. Initial image of Yosemite sequence.

We evaluated the Euclidean distance between our match results and ground truth. Figure 3 shows the distribution of the pixel match errors over the image. White regions indicate that the pixel match errors are below half a pixel. The next darker level indicates pixel errors of between half and one pixel. The meaning of the other shades of gray are given in the legend of figure 3. Note that since we try to match pixel onto pixel, half a pixel error is as good as we can statistically expect the result to be. Large areas have been matched very well, whereas there are problems in the area of the mountain on the left. Nevertheless, problematic areas are locally confined, which shows the robustness of the algorithm. Recall that we only used two images to evaluate the optic flow. By extending the algorithm to incorporate more images of a flow sequence we hope to improve the matching quality further.

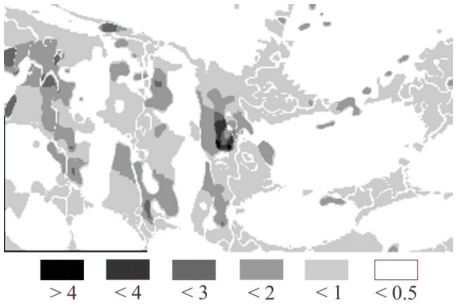


Fig. 3. Distribution of matching errors.

Although the algorithm has a simple mathematical structure, its computational complexity is high. Nevertheless, in principle the match likelihood estimation of all pixels can be done in parallel. In fact, each element of the pixel match probability distributions can be regarded as a single neuron which performs a

simple calculation. Evaluating each neuron is all that has to be done per iteration. We have implemented a similar structure on an FPGA which shows good preliminary results.

Of course, there are still a number of problems that have to be addressed by future research. Nevertheless, the results obtained with the algorithm show that despite its simple structure, it is a good dense image point matcher. Note that a program called *Acre* to test the algorithm on arbitrary images, is available from the web page of the first author (`www.perwass.de`).

References

[1] L. Florack, W. Niessen, and M. Nielsen. The intrinsic structure of optic flow incorporating measurment duality. *International Journal of Computer Vision*, 27(3):263–286, 1998.

[2] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.

[3] G. le Besnerais and H. Oriot. Disparity estimation for high resolution stereoscopic reconstruction using the gnc approach. In *Proc. IEEE Int. Conf. on Image Processing*, volume 2, pages 594–597, 1998.

[4] C.L. Zitnick and T. Kanade. A cooperative algorithm for stereo matching and occlusion detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):675–684, 2000.

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximare energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[6] Peter N. Belhumeur. A Bayesian approach to binocular stereopsis. *International Journal of Computer Vision*, 19:237–262, 1996.

[7] M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29:5–28, 1998.

[8] N. Vasconcelos and A. Lippman. Empirical bayesian motion segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):217–221, 2001.

[9] P.H.S. Torr, R. Szeliski, and P. Anandan. An integrated bayesian approach to layer extraction from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):297–303, 2001.

[10] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[11] J.L. Marroquin, F.A. Velasco, M. Rivera, and M. Nakamura. Gauss-markov measure field models for low-level vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):337–348, 2001.

[12] D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194:209–236, 1976.

[13] C.B.U. Perwass and G. Sommer. A fuzzy logic algorithm for dense image point matching. In *Proceedings of Vision Interface 2001*, pages 39–47, 2001.

[14] C.B.U. Perwass and G. Sommer. Dense image point matching through propagation of local constraints. Bericht Nr. 0205, Christian-Albrechts-Universität Kiel, 2002.

[15] D. Scharstein and R. Szeliski. Stereo matching with nonlinear diffusion. *International Journal of Computer Vision*, 28(2):155–174, 1998.

Fast Phase-based Orientation Estimation for Panoramic Images

Wolfgang Stürzl and Hanspeter A. Mallot

Universität Tübingen, Zoologisches Institut, Kognitive Neurowissenschaften,
72076 Tübingen, Germany
wolfgang.stuerzl@uni-tuebingen.de
WWW institute home page: http://www.uni-tuebingen.de/cog/

Abstract. We propose an algorithm for the fast estimation of the shift that maximizes correlation of panoramic images. By using Fourier transformation we achieve sub-pixel accuracy. We use a coarse-to-fine approach exploiting the fact that in natural images low spatial frequencies have higher spectral power. Starting from maximization of image correlation, we derive an expression for the estimated shift, consisting of a weighted sum of frequency-dependent shifts. After estimation of the optimal shift, correlation has to be computed only once for each (panoramic) image comparison. This reduces the complexity compared to the standard approach where all orientations have to be tested from $O(N^2)$ to $O(N)$ (N: number of pixels per row). We also introduce a scalar measure of local image variation where we use the fast shift estimation to find the optimal orientation of neighboring images.

1 Introduction

In recent years numerous articles have been published about the use of panoramic images for robot localization, e.g. [1]. In a purely image-based approach where no additional information about relative orientation is available, the correlation of two panoramic images consisting of N pixels[1] (image vectors $\mathbf{I},\mathbf{J}$, $\mathbf{I} := (I_0, I_1, \ldots, I_{N-1})^T$) usually has to be performed by testing all possible orientations (see e.g. [2]), i.e.[2]

$$\mathrm{corr}(\mathbf{I},\mathbf{J}) := \max_s \mathrm{corr}(\mathbf{I},\mathbf{J},s) \tag{1}$$

$$\mathrm{corr}(\mathbf{I},\mathbf{J},s) := \sum_x I_{[(x+s) \bmod N]} J_x \tag{2}$$

This is time consuming, if a large image data base has to be searched. Therefore a method for the fast estimation of relative orientation is desirable.

With this paper we propose an algorithm for the fast sub-pixel estimation of the shift that maximizes correlation using Fourier transformation of panoramic images.

[1] Although we consider only one dimensional panoramic images in this paper the extension to two dimensions is fairly straightforward.

[2] To simplify notation we assume that image vectors have zero mean and unit length.

2 Fourier representation and shift estimation

The correlation of two images with complex Fourier coefficients $\{C_k = |C_k|e^{i\varphi_k}\}$, $\{D_k = |D_k|e^{i\psi_k}\}$ can be rewritten using $C_0 = D_0 = 0$ and $C_{-k} = C_k^*$ (since pixels are real values),

$$\begin{aligned}
\mathrm{corr}(\mathbf{I},\mathbf{J},s) &= \sum_x I_{[(x+s) \bmod N]} J_x && (3)\\
&= \sum_{x=0}^{N-1} \sum_{l=-N/2}^{N/2} C_l e^{-i\frac{2\pi}{N}l(x+s)} \sum_{k=-N/2}^{N/2} D_k e^{-i\frac{2\pi}{N}kx} && (4)\\
&= N \sum_{k=1}^{N/2} 2\,\mathrm{Re}[e^{i\frac{2\pi}{N}ks} C_k^* D_k] && (5)\\
&= 2N \sum_{k=1}^{N/2} |C_k||D_k| \cos(\varphi_k - \psi_k - \omega_k s)\ , \quad \omega_k := \frac{2\pi}{N}k\ . && (6)
\end{aligned}$$

Searching for s that maximizes the correlation, we compute the derivative:

$$\begin{aligned}
0 &= \frac{\partial}{\partial s}\mathrm{corr}(\mathbf{I},\mathbf{J},s) \quad \Longleftrightarrow && (7)\\
0 &= \sum_k |C_k||D_k| \sin(\varphi_k - \psi_k - \omega_k s)\omega_k && (8)
\end{aligned}$$

If $\mathbf{I}$ and $\mathbf{J}$ are approximately shifted versions of each other $|C_k| \approx |D_k|$ holds and there exists a shift $s \in [0, N[$ and integers n_k which satisfy

$$\begin{aligned}
\psi_k &\approx \varphi_k - \omega_k s + 2\pi n_k, \quad \forall k && (9)\\
\Longleftrightarrow \quad 0 &\approx \varphi_k - \psi_k - \omega_k s + 2\pi n_k && (10)\\
\Longleftrightarrow \quad s \approx s_k &:= \frac{\varphi_k - \psi_k + 2\pi n_k}{\omega_k} = \frac{N}{k}\Big(\frac{\varphi_k - \psi_k}{2\pi} + n_k\Big)\ . && (11)
\end{aligned}$$

Hence, if n_k is unknown, only for $k = 1$ we get an unique $s_k \in [0, N[$. Using (10) we approximate

$$\begin{aligned}
\sin(\varphi_k - \psi_k - \omega_k s) &= \sin(\varphi_k - \psi_k - \omega_k s + 2\pi n_k) && (12)\\
&\approx \varphi_k - \psi_k - \omega_k s + 2\pi n_k && (13)
\end{aligned}$$

and obtain from (8),

$$\begin{aligned}
0 &\approx \sum_k |C_k||D_k|\omega_k(\varphi_k - \psi_k - \omega_k s + 2\pi n_k) && (14)\\
\Longleftrightarrow \quad s \approx \bar{s} &:= \frac{\sum_k |C_k||D_k|\omega_k(\varphi_k - \psi_k + 2\pi n_k)}{\sum_k |C_k||D_k|\omega_k^2} && (15)\\
&\overset{(11)}{=} \frac{\sum_k |C_k||D_k|\omega_k^2 s_k}{\sum_k |C_k||D_k|\omega_k^2} = \frac{\sum_k \alpha_k s_k}{\sum_k \alpha_k}\ , && (16)
\end{aligned}$$

where we have defined $\alpha_k := |C_k||D_k|\omega_k^2$.

Considering (11), (16) and the fact that the integers n_k are unknown, we propose the following coarse-to-fine algorithm for the estimation of shift s:

$$\bar{s}_1 = s_1 = \Big(\frac{\varphi_1 - \psi_1}{\omega_1}\Big) \bmod N \qquad \alpha_1 = |C_1||D_1|\omega_1^2 \tag{17}$$

For $k = 2, 3, \ldots, K\ (\leq N/2)$ do:

$$n_k = \mathrm{rint}\Big[\frac{\psi_k - \varphi_k + \omega_k \bar{s}_{k-1}}{2\pi}\Big] \quad \in \mathbb{Z} \tag{18}$$

$$s_k = \frac{\varphi_k - \psi_k + 2\pi n_k}{\omega_k} \qquad \alpha_k = |C_k||D_k|\omega_k^2 \tag{19}$$

$$\bar{s}_k = \frac{\sum_{l=1}^{k} \alpha_l s_l}{\sum_{l=1}^{k} \alpha_l} = \frac{(\sum_{l=1}^{k-1} \alpha_l)\bar{s}_{k-1} + \alpha_k s_k}{\sum_{l=1}^{k-1} \alpha_l + \alpha_k} \quad , \tag{20}$$

where 'rint[ ]' means "round to the nearest integer".
Using the estimated shift $s^{\mathrm{est}} := \bar{s}_K \bmod N$, the correlation according to (6) can be computed.

The algorithm reduces the complexity from $O(N^2)$ to $O(N)$. Because of (18) the correct estimation of $\bar{s}_1$ is crucial for the coarse-to-fine procedure. Although occurring rarely (since natural images usually have high spectral power in low frequencies), small and hence noisy values of $|C_1|$ and $|D_1|$ are therefore critical. In this case (α_1 small) we compute a second estimation of s starting with $\bar{s}_1' = (\bar{s}_1 + N/2) \bmod N$. The shift with the larger correlation is then assumed to be the correct estimation. (If successive α_l, $l = 1, 2, \ldots$, are small, even more possible estimations of s will have to be considered, but this case never occurred for the data base used in Sect. 3). However single small values of α_k, $k > 1$ are not critical since $\bar{s}_k$ is calculated as a weighted sum.

Computation of Fourier coefficients can be done during the exploration/image acquisition phase to speed up shift estimation and image correlation. The number K of Fourier coefficients that have to be used for the calculation depend on the required accuracy and the frequency distribution of the noise.

3 Comparison to "standard correlation", equation (1)

Panoramic images were recorded during a full rotation of a Khepera robot and correlated with the image taken at 0° orientation. Rotation angles were measured using the robot's odometry. Panoramic vision is achieved by a CCD-camera mounted on top of the Khepera and directed vertically towards a conic mirror, as described in [2]. Panoramic 1D-images consisting of $N = 72$ pixels, each of them representing a 5° region (vertically) around the horizon, were extracted from the camera images. The Khepera did the full rotation in steps of approximately 0.5° until image similarity reaches a maximum. To correct small odometric errors, the measured values were thereafter scaled to give the full range of 360°. The drift during rotation was within the accuracy of the tracking system (≈ 2 mm).

Due to an occlusion range of 3 pixels (= 15°) caused by the cable for video transmission to the host computer, images taken at different rotation angles do

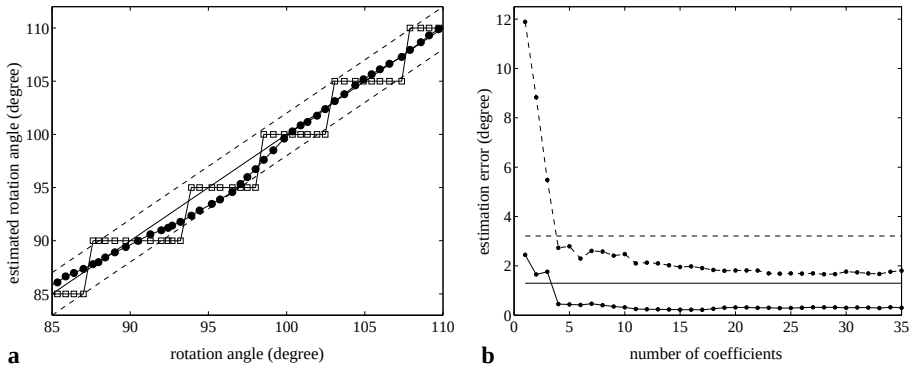


Fig. 1. a: Estimation of rotation angle using the phase-based calculation (filled circles) with 15 Fourier coefficients compared to standard correlation (rectangles). The straight continous line shows the corrected angle measured by odometry ϕ^{odo} (see text), dashed lines represent $\phi^{\text{odo}} \pm 2°$. Only the range [85°, 110°] is shown where the largest error of approximately 2° occurs. **b:** Maximum (dashed curve) and mean (continuous curve) of the estimation error of the rotation angle, i.e. $|\phi^{\text{odo}} - \phi^{\text{est}}|$, in dependence of the number of Fourier coefficients used for the calculation (horizontal lines show the corresponding errors using standard correlation).

not overlap completely (pixel values in the occluded range were linearly interpolated from the neighboring pixel values). Hence small differences of the Fourier amplitudes and phases occur and lead to errors in the estimated rotation angle.

While standard correlation allows calculation of the rotation angle only in steps of 5°, the proposed algorithm achieves sub-pixel estimations (Fig. 1 a). As can be seen in Fig. 1 b, the use of only four coefficients results in a smaller orientation error (mean $\approx 0.45°$, max $\approx 2.73°$) compared to the standard correlation (mean $\approx 1.30°$, max $\approx 3.21°$).

4 Calculation of local image variation and comparison to "zero phase representation"

Since the algorithm is intended to be used for image-based localization, we investigate the performance in the vicinity of places where images were recorded. We introduce a (scalar) measure of local image variation,

$$\text{liv}(\mathbf{x}) := \sqrt{\left(\frac{\partial \mathbf{I}(\mathbf{x})}{\partial x}\right)^2 \left(\frac{\partial \mathbf{I}(\mathbf{x})}{\partial y}\right)^2 - \left(\frac{\partial \mathbf{I}(\mathbf{x})}{\partial x}\frac{\partial \mathbf{I}(\mathbf{x})}{\partial y}\right)^2} = \sqrt{\det \hat{g}(\mathbf{x})} \qquad (21)$$

$$\hat{g}(\mathbf{x}) := \begin{pmatrix} \left(\frac{\partial \mathbf{I}(\mathbf{x})}{\partial x}\right)^2 & \frac{\partial \mathbf{I}(\mathbf{x})}{\partial x}\frac{\partial \mathbf{I}(\mathbf{x})}{\partial y} \\ \frac{\partial \mathbf{I}(\mathbf{x})}{\partial x}\frac{\partial \mathbf{I}(\mathbf{x})}{\partial y} & \left(\frac{\partial \mathbf{I}(\mathbf{x})}{\partial y}\right)^2 \end{pmatrix} , \qquad (22)$$

where $\mathbf{x} = (x, y) \in \mathbb{R}^2$ is a parameterization of the image manifold $\mathbf{I}(x, y)$. $\hat{g}$ is known as metric tensor in differential geometry. A surface element in the image

space corresponding to a small area $D \subset \mathbb{R}^2$ is given as

$$S(D) := \iint_D \mathrm{liv}(x,y)\,dxdy \approx \mathrm{liv}(\mathbf{x}_S) \iint_D dxdy = \mathrm{liv}(\mathbf{x}_S)D \tag{23}$$

$$\Longrightarrow \quad \mathrm{liv}(\mathbf{x}_S) \approx \frac{S(D)}{D} \quad , \tag{24}$$

where $\mathbf{x}_S$ is the centroid of D. Three image vectors at neighboring points $\mathbf{I}(\mathbf{x}_1)$, $\mathbf{I}(\mathbf{x}_2)$, $\mathbf{I}(\mathbf{x}_3)$ define a triangle in image space,

$$S_\triangle := \frac{1}{2}|\Delta\mathbf{I}_{12}||\Delta\mathbf{I}_{13}||\sin\angle(\Delta\mathbf{I}_{12},\Delta\mathbf{I}_{13})| \tag{25}$$

$$= \frac{1}{2}\sqrt{\Delta\mathbf{I}_{12}^2\Delta\mathbf{I}_{13}^2 - (\Delta\mathbf{I}_{12}\Delta\mathbf{I}_{13})^2} \quad , \quad \Delta\mathbf{I}_{1a} := \mathbf{I}(\mathbf{x}_a) - \mathbf{I}(\mathbf{x}_1),\ a = 2,3 \ . \tag{26}$$

The corresponding triangle in $\mathbb{R}^2$ is

$$D_\triangle := \frac{1}{2}|\Delta\mathbf{x}_{12}||\Delta\mathbf{x}_{13}||\sin\angle(\Delta\mathbf{x}_{12},\Delta\mathbf{x}_{13})| \tag{27}$$

$$= \frac{1}{2}\sqrt{\Delta\mathbf{x}_{12}^2\Delta\mathbf{x}_{13}^2 - (\Delta\mathbf{x}_{12}\Delta\mathbf{x}_{13})^2} \quad , \quad \Delta\mathbf{x}_{1a} := \mathbf{x}_a - \mathbf{x}_1,\ a = 2,3 \ . \tag{28}$$

Hence we can approximate

$$\mathrm{liv}(\mathbf{x}_S) \approx \frac{S_\triangle}{D_\triangle} = \sqrt{\frac{\Delta\mathbf{I}_{12}^2\Delta\mathbf{I}_{13}^2 - (\Delta\mathbf{I}_{12}\Delta\mathbf{I}_{13})^2}{\Delta\mathbf{x}_{12}^2\Delta\mathbf{x}_{13}^2 - (\Delta\mathbf{x}_{12}\Delta\mathbf{x}_{13})^2}} \quad , \tag{29}$$

$$\mathbf{x}_S = \frac{1}{3}(\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3) \ . \tag{30}$$

We evaluated a data base consisting of 1250 panoramic images, which were automatically recorded by a small Khepera robot (width ≈ 5 cm, height ≈ 13 cm) inside a toy house arena of approximate size 140 cm $\times$ 120 cm. Recording positions were approximately on a rectangular 44×36 grid with cell size 2.5×2.5 cm^2. Minimum distances of recording positions to the arena walls were ≈ 15 cm, minimum distances to toy houses were ≈ 5 cm. The accuracy of the tracking system used for the estimation of the robot's pose is ≈ 2 mm (position) and $1.5°$ (orientation). We calculated the average power spectrum $|P(f)|^2$ of the images in the data base as described in Chap. 2 of [3]. The resulting exponent of -1.84 ± 0.24, i.e. $|P(f)|^2 \propto 1/f^{1.84}$, is in the range reported by other studies (summarized in [3]) and is in agreement with the empirical finding that in most natural images low frequencies are dominant.

Using three image vectors at points of an approximately rectangular triangle, local image variation was calculated as follows: After choosing one reference image $\mathbf{I}_1$, images $\mathbf{I}_2$ and $\mathbf{I}_3$ were rotated by shifts which maximize $\mathrm{corr}(\mathbf{I}_1,\mathbf{I}_2)$ and $\mathrm{corr}(\mathbf{I}_1,\mathbf{I}_3)$ using (20) for the shift estimations and inverse Fourier transformation. Then liv and $\mathbf{x}_S$ according to (29) and (30) were calculated. Each grid cell consisting of four recording positions was divided into two triangles resulting in approx. 3000 liv-values for the whole image data base.

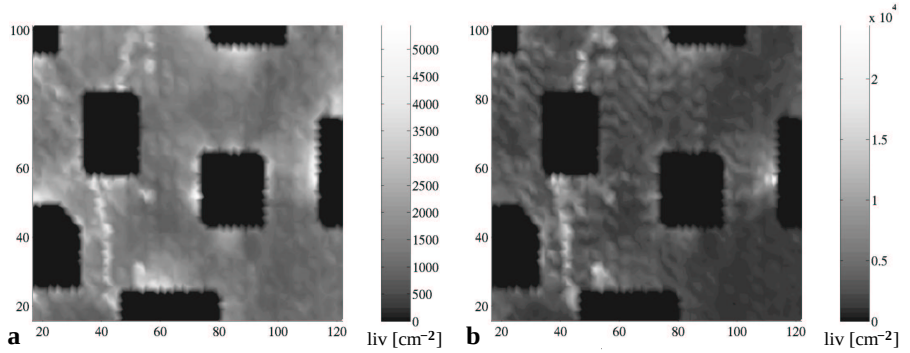


Fig. 2. a: Local image variation calculated using the image data base and (20) with 25 Fourier coefficients to optimize orientation of images. Bright areas have high liv-values (see gray scale). Continuous dark regions depict toy houses where no images were recorded. **b:** Calculated liv-values using the zero phase representation of panoramic images. Since at several positions values are much higher than in **a**, relative orientation differences of proximate images were not sufficiently reduced at these locations.

In Fig. 2 a calculated liv-values in the toy house arena are shown as gray values. As expected, high values occur at positions near objects (toy houses) with texture of high contrast. We expect that locations with high liv-values allow better localization than places with low values, i.e. liv-value can be used as a *local* heuristic measure where snapshots should be taken in order to achieve efficient image-based localization. This will be investigated in future work.

In [4] the *zero phase representation* for panoramic images was suggested yielding orientation-independent representations: Panoramic images are rotated in order to obtain phase $\varphi_1 = 0$ for the first Fourier coefficient. To perform image-based localization, image correlation has to be computed only once (in the resulting orientation) since images at proximate positions are expected to receive approximately the same orientation. However, as can be seen in Fig. 2 b, at several locations in the toy house arena, the use of only one Fourier phase is not enough to reduce the image shift of neighboring images sufficiently, resulting in significantly higher liv-values compared to Fig. 2 a. At these places image-based localization using the zero phase representation is likely to fail.

References

1. Jogan, M., Leonardis, A.: Robust localization using eigenspace of spinning-images. In: Workshop on Omnidirectional Vision, IEEE Computer Society (2000) 37–44
2. Franz, M.O., Schölkopf, B., Mallot, H.A., Bülthoff, H.H.: Where did I take that snapshot? Scene-based homing by image matching. Biol. Cybern. **79** (1998) 191–202
3. van der Schaaf, A.: Natural Image Statistics and Visual Processing. PhD thesis, Delft University of Technology (1998)
4. Pajdla, T., Hlavac, V.: Zero phase representation of panoramic images for image based localization. In Solina, F., Leonardis, A., eds.: International Conference on Computer Analysis of Images and Patterns, Springer Verlag (1999) 550–557

Multimodal integration

Polymodal space representation in primate posterior parietal cortex (PPC)

Frank Bremmer[1,2,3], Anja Schlack[1], Gereon R. Fink[2], and Klaus-Peter Hoffmann[1]

[1] Allg. Zoologie & Neurobiologie, Ruhr-Universität Bochum, D-44780 Bochum, Germany
{schlack, kph}@neurobiologie.ruhr-uni-bochum.de
[2] Institut für Medizin, Forschungszentrum Jülich, D-52425 Jülich, Germany
g.fink@fz-juelich.de
[3] New address: Neurophysik, Philipps-Universität Marburg, D-35032 Marburg, Germany
frank.bremmer@physik.uni-marburg.de

Abstract. The primate posterior parietal cortex (PPC) is involved in the multisensory processing of spatial information. Damage to this part of the cerebrum leads to marked, and often long lasting, disturbances in spatial perception and visually guided action. Much has been learned about the underlying cortical mechanism subserving spatially oriented behavior in the last twenty years, due in large part to the development of awake primate behavioral physiology, to detailed investigations of behavioral deficits following brain damage in humans and to functional imaging of normal human volunteers. This review aims at describing some of the underlying neuronal circuits involved in spatial processing as has been revealed by single cell recordings in awake monkeys and fMRI studies in healthy human subjects. Both approaches, used in parallel, have led to an improved understanding of the basic principles of the processing of spatial information in the primate brain.

1 Introduction

The primate PPC is related to the processing of spatial and motion information [1]. In humans, damage to this *where* or *how* pathway, in particular in the right hemisphere, leads to behavioral deficits often referred to as *extinction* and *neglect* [2]. While extinction describes the inability to perceive a contralateral stimulus that is presented simultaneously with an ipsilateral one, neglect refers to the inability of perceiving (objects in) the contralateral space in general. Two specific functional aspects of neglect (and/or extinction) are essential for the description of this behavioral deficit and might be crucial for the understanding of how normal posterior parietal cortex operates. Firstly, these patients sometimes look at points in space contralateral to their lesion site although they do not perceive what is there (see e.g. [3]). Accordingly, some of these spatial locations that are not perceived, are located ipsilaterally with respect to the fovea but contralaterally with respect to the head or the body. This implies that the observed behavioral deficit occurs not in an eye- but rather in a head- or body-centered frame of reference. Secondly, extinction and neglect can occur across different sensory modalities, i.e. they are polymodal [4].

Lesions of posterior parietal and frontal cortex lead to comparable behavioral deficits in humans and non-human primates. It therefore appears appropriate to (i) consider the macaque monkey as an animal model for the better understanding of the normally working posterior parietal cortex and (ii) test for functional equivalencies between humans and macaques concerning specific cortical regions which have been described in detail for the macaque.

2 Polymodal motion responses in macaque PPC

Recent neurophysiological studies in macaque monkeys revealed a number of functionally distinct subdivisions along and within the intraparietal sulcus (IPS). One of these areas is the ventral intraparietal area (VIP) located in the fundus of the IPS. Based on anatomical data, area VIP was originally defined as the MT projection zone in the intraparietal sulcus (IPS) [5]. This anatomical result suggested that neurons in area VIP might be responsive for the direction and speed of moving visual stimuli, and in general might encode self-motion information.

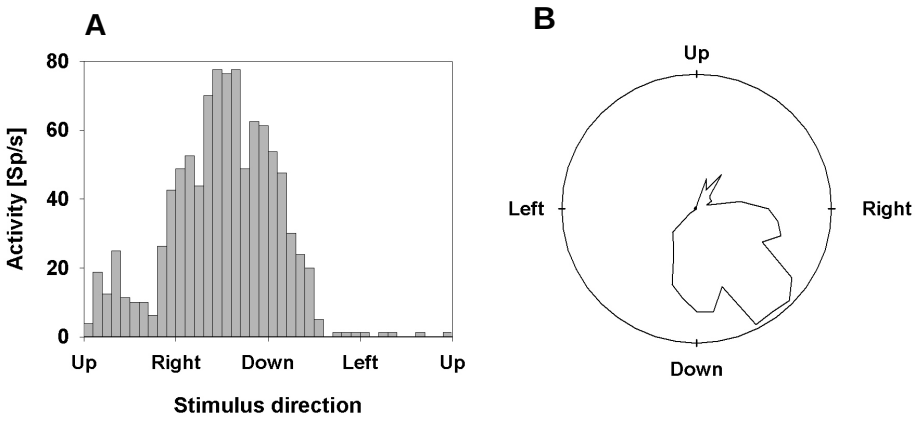


Fig. 1. Response of a neuron from area VIP to a visual stimulus moving on a circular pathway. Response is shown as histogram in (A) and as polar plot (B). This neuron clearly prefers stimulus motion right- and downward.

Recent studies confirmed that VIP neurons respond selectively to basic optic flow pattern like frontoparallel motion, or forward or backward motion [6,7]. An example for responsiveness to frontoparallel motion is shown in Figure 1. Both panels show the response of a VIP neuron to movement on a circular pathway. Data are shown as response histogram (A) and in a polar plot (B). It is obvious that this neuron responds best to stimulus motion down and to the right. Figure 2 shows the response of a cell preferring visually simulated forward over backward motion. Responses are shown for an expansion stimulus simulating forward motion (A) and a contraction stimulus

simulating backward motion (B). A total of 70% of the cells in area VIP responds selectively to frontoparallel motion and/or forward or backward motion. Interestingly, preference for forward or backward motion cannot be predicted by knowledge of the location of a neuron's visual receptive field and its preference for frontoparallel motion. It therefore appears that the responsiveness for forward or backward motion is an inherent response property of cells in area VIP.

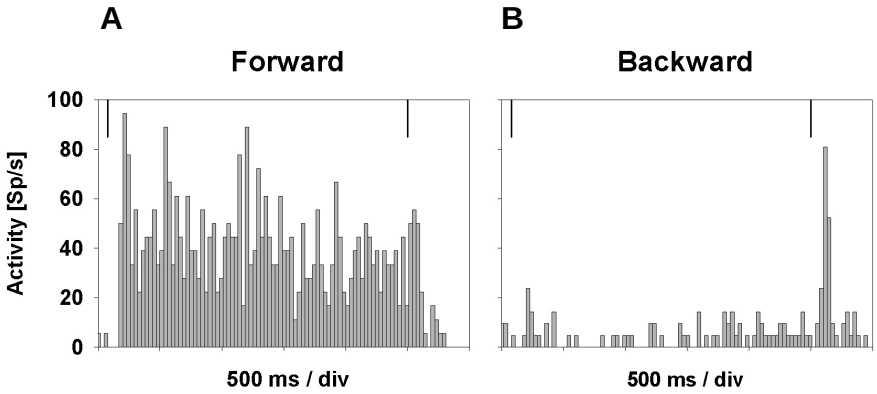


Fig. 2. Response to visually simulated forward and backward motion. The two histograms show the response of a neuron from area VIP to an expansion stimulus simulating forward motion (A) and a contraction stimulus simulating backward motion (B). Vertical lines within the panels indicate the onset and offset of motion. Response differences were statistically significant at $p<0.001$.

Like visual information, somatosensory signals can be used to encode motion information. Many neurons in area VIP respond also to tactile stimulation [8,9]. Most VIP cells that have a somatosensory receptive field (RF) show a positive response to passive superficial stimulation of restricted portions of the head, with the upper and lower face areas being represented equally often. Tactile and visual RFs are organized in an orderly manner with tactile RFs showing a systematic relation to the main axes of the visual field. Critically, the matched tactile and visual RFs often demonstrate co-aligned direction selectivity.

Another source of motion information may result from vestibular stimulation, i.e. rotational and/or translational self-motion. Accordingly, neurons in area VIP were tested for their responses to vestibular stimulation. About one third of the neurons respond with direction selective discharges during whole-body sinusoidal horizontal rotational movement [10]. All neurons with rotational vestibular responses also show directionally selective visual responses. Interestingly, preferred directions for visual and for rotational vestibular stimulation are co-directional, i.e. non-synergistic, or non-complementary.

These response characteristics led us to hypothesize that area VIP might be involved in the encoding of visual motion in near extrapersonal space. We thus tested neurons for their sensitivity to horizontal disparity. Random dot patterns moving along a circular pathway were presented at one of seven disparities, ranging from -3° (near) to 3° (far) disparity. These disparity values correspond to stimuli located

between 27cm (-3°) and 223cm (3°) in front of the monkey. An example for the responses of a cell with a maximum discharge for the nearest stimulus is shown in Figure 3A. The population histogram on the right (B) indicates that this response characteristic was quite common in area VIP and significantly different from a uniform distribution (chi^2-test: $p<0.001$): 70% of the neurons had their response maximum for stimuli in near space while only 21% of the neurons preferred stimuli presented in far space. The remaining cells had their response maximum within the plane of fixation. Our data therefore supply evidence for the proposed role of area VIP for the encoding of motion in near extra- personal space.

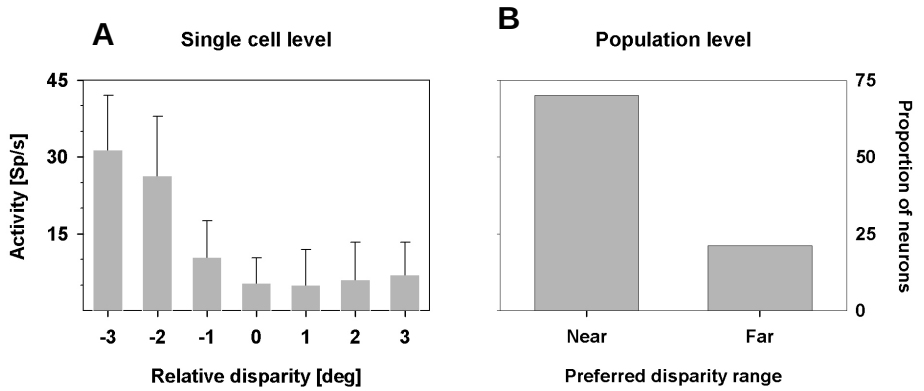


Fig. 3. Disparity selectivity in area VIP. Panel (A) shows the response of a single neuron for visual stimuli presented at different (virtual) depths. Bars indicate the mean response (+ sd) for a random dot pattern moving into the cell's preferred direction. Response is strongest for the nearest stimulus ($p<0.001$) presented at -3° disparity. Panel (B) indicates that this preference for near stimuli was a common response behavior for the population of cells (n=90) tested.

Sensory signals arising from different modalities are encoded in different frames of reference. While vestibular signals and tactile information (arising from stimulation of RFs on the head) are encoded in craniocentric coordinates, visual information is initially encoded retinocentrically. This led to the question, whether information from different sensory modalities might be encoded in a common, probably craniocentric reference frame. Accordingly, area VIP was tested for the existence of head-centered cells by measuring the location of visual RFs for different fixation locations [11]. A wide range of RF types was found. Some neurons had an RF that moved rigidly with the eyes, while other neurons encoded the same location in space irrespective of eye position. Such cells code visual information in a head-centered frame of reference. Interestingly enough, many cells had intermediate reference frames: they compensated only in part for the underlying gaze shift. While it was initially unclear whether these intermediate encoding cells represented an incomplete computational step from an eye-centered to a head-centered representation, there is now evidence from computational studies that these intermediate types arise naturally in neural networks involved in polymodal space representation [12].

3 Polymodal motion responses in human PPC

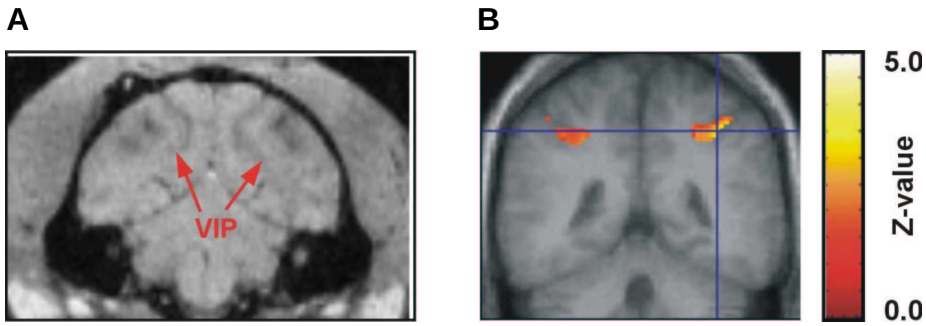


Fig. 4. Localization of area VIP in the macaque (A) and the human (B) posterior parietal cortex. The left panel (A) shows an anatomical MRI from one of the monkeys involved in the experiments. The right panel (B) shows a superposition of an anatomical MRI from the group of subjects (n=8) involved in the study and the region within the PPC activated by polymodal (visual, tactile, auditory) motion signals.

From the aforementioned, it becomes obvious that the question needed to be explored, whether or not in humans an equivalent area to macaque area VIP exists. Accordingly, the test for the existence of 'human area VIP' was based on one of its most prominent response features in the macaque, i.e. sensory responses to polymodal motion stimuli. In this functional MRI study, subjects experienced a visual (large random dot pattern), tactile (air flow) or auditory (binaural beats) motion stimulus or a stationary control. Significant cortical activation ($p<0.05$, corrected) was observed for each individual stimulus condition. Conjunction analysis revealed cortical structures activated by motion in all three modalities, i.e. vision, touch, and audition. Bilateral significant activation was found in three circumscribed cortical regions, one of which was located in the PPC. By superimposing the functional images on the average anatomical brain originating from the group of subjects (Figure 4) it was possible to identify the activated region as lying in the depth of the IPS [13]. Based on these functional and anatomical characteristics it was suggested to consider this area to be the functional equivalent of macaque area VIP.

4 Conclusion

Complementary studies of macaque single cell recordings and fMRI in humans helped to elucidate the functional role of the PPC in polymodal spatial perception and motion encoding. In addition, the reviewed findings relate to neuropsychological deficits observed in patients with (most often right) posterior parietal lobe injury: The most prominent functional features of macaque area VIP are (i) responses to polymodal stimuli predominantly in near extrapersonal space, and (ii) the encoding of sensory information from different modalities in a head- or body-centered frame of

reference [11]. It is exactly this related type of attentive behavioral sensorimotor deficit, which in patients most often results from lesions centered on the (right) PPC.

5 Acknowledgements

This work was supported by grants from the *Deutsche Forschungsgemeinschaft.*

6 References

1. Ungerleider,L.G. & Mishkin,M. Analysis of visual behavior. Ingle,D.J., Goodale,M.A. & Mansfield,R.J.W. (eds.), pp. 549-586 (MIT Press, Cambridge, MA,1982).
2. Driver,J. & Mattingley,J.B. Parietal neglect and visual awareness . *Nat. Neurosci.* 1, 17-22 (1998).
3. Husain,M. *et al.* Impaired spatial working memory across saccades contributes to abnormal search in parietal neglect. *Brain* 124, 941-952 (2001).
4. Ladavas,E., Di Pellegrino,G., Farne,A. & Zeloni,G. Neuropsychological evidence of an integrated visuotactile representation of peripersonal space in humans. *J. Cogn. Neurosci.* 10, 581-589 (1998).
5. Maunsell,J.H.R. & Van Essen,D.C. The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *J. Neurosci.* 3 , 2563-2580 (1983).
6. Schaafsma,S.J. & Duysens,J. Neurons in the ventral intraparietal area of awake macaque monkey closely resemble neurons in the dorsal part of the medial superior temporal area in their responses to optic flow patterns. *J. Neurophysiol.* 76, 4056-4068 (1996).
7. Bremmer,F., Duhamel,J.-R., Ben Hamed,S. & Graf,W. Heading encoding in the macaque ventral intraparietal area (VIP). Eur.J.Neurosci . 2002 (In Press)
8. Colby,C.L., Duhamel,J.-R. & Goldberg,M.E. Ventral intraparietal Area of the macaque: anatomical location and visual response properties. *J. Neurophysiol.* 69, 902-914 (1993).
9. Duhamel,J.R., Colby,C.L. & Goldberg,M.E. Ventral intraparietal area of the macaque: congruent visual and somatic response properties. *J. Neurophysiol.* 79, 126-136 (1998).
10. Bremmer,F., Klam,F., Duhamel,J.-R., Ben Hamed,S. & Graf,W. Visual-vestibular interactive responses in the macaque ventral intraparietal area (VIP). Eur.J.Neurosci . 2002 (In Press)
11. Duhamel,J.R., Bremmer,F., BenHamed,S. & Graf,W. Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature* 389, 845-848 (1997).
12. Deneve,S., Latham,P.E. & Pouget,A. Efficient computation and cue integration with noisy population codes. *Nat. Neurosci* 4, 826-831 (2001).
13. Bremmer,F. *et al.* Polymodal motion processing in posterior parietal and premotor cortex: a human fMRI study strongly implies equivalencies between humans and monkeys. *Neuron* 29, 287-296 (2001).

Intersensory Interaction in Arm and Eye Movements

Petra A. Arndt

Carl von Ossietzky Universität Oldenburg, Institut für Kognitionsforschung,
26111 Oldenburg, Germany
petra.arndt@uni-oldenburg.de

Abstract. To investigate whether eye movements and arm movements share motor control processes or are programmed separately we analysed the characteristics of multisensory, visual-auditory integration in eye and arm movements using a focussed attention paradigm. The effects of spatio-temporal visual-auditory stimulus relationship, found in a first experiment, contradict the notion of common control processes. In contrast, no evidence for separate movement programming was found in a second experiment with variation of auditory stimulus intensity. These conflicting results indicate that brain structures in charge of hand movement control may have the capability of a higher spatial resolution for auditory stimuli. A third experiment gives an indication of the origin of the higher spatial resolution and supports the notion of a common visual-auditory representation as a basis for eye and arm movement control.

1 Introduction

The question of whether saccadic eye movements and goal directed arm movements share common processing stages or are programmed separately is still under debate. Recent physiological findings have provided new evidence for a combined representation of eye and arm movements in several brain areas [1,2].

Three experiments are presented which investigate visually guided eye and arm movements under visual-auditory stimulation. We employed a focussed attention paradigm where subjects are asked to respond to the visual target stimulus and to ignore an accessory auditory stimulus. However, although the auditory stimulus is to be ignored it has specific effects on the performance of movements [3,4]. These effects change with the variation of temporal and spatial relationship between visual and auditory stimulus. Given that eye and arm movements share processes based on the same multimodal representation of sensory stimuli, the effects of the auditory stimulus in dependence of spatiotemporal stimulus arrangement should be the same for both movements.

The neutral basis for this is that both multimodal, visual-auditory neurons as well as arm-movement-related neurons were found in certain brain structures, e.g. the superior colliculus [5].

2 General Methods

2.3 Participants

Ten paid volunteers with normal or corrected to normal visual acuity took part in the experiment. Participants reported to have no hearing problems of any kind. All participants were naïve towards the purpose of the study.

2.2 Apparatus and Stimuli

Experiments 1 and 2 were performed in a virtual auditory environment, whereas we used a free-field setup with loudspeakers for Experiment 3.

Participants were seated in a dark sound proof room with the head supported by a table-mounted chin rest. Unless arm movements were executed both forearms were resting on the table.

Auditory stimuli. White noise signals (band-passed 0.2 to 20 kHz, 500 *ms*) were used as stimuli in all experiments. For virtual acoustics these signals were convolved with the head related transfer function of a dummy head and played back via a high precision sound card on headphones. Noise signals for free-field stimulation, generated by a TDT System (Tucker Davis Technologies), were displayed by two loudspeakers placed at the subject's eye level.

Visual stimuli. In Experiments 1 and 2, white dots (diameter 0.1°) presented on a black monitor screen (37") were used as central fixation point and peripheral visual target stimuli. In Experiment 3, two red light emitting diodes attached to the loudspeakers served as visual stimuli; a third, central LED as fixation point.

The respective spatio-temporal stimulus relationships, stimulus onset asynchronies (SOAs) and intensities are given in sections 3 to 5.

Data recording. Eye movements were measured with an infrared light reflecting system (IRIS, Skalar Medicals) providing an analog signal of horizontal eye position and velocity. Data were recorded with a sampling rate of 1 kHz. In Experiment 1 and 2 a joystick placed midline in front of the participant was used to measure goal directed arm movements. In the third experiment a photoelectric switch was used to collect reaction time data of the arm movements and a magnetic position tracker (Polhemus Frastrack) to register movement trajectories.

2.3 Procedure

Each trial started with the presentation of the fixation point for a random time interval of between 800 and 2000 *ms*. The visual stimuli were presented to the left or right in pseudorandom order after extinction of the fixation point for 500 *ms*, either alone or with an auditory stimulus. The spatial distance between visual and auditory stimulus and SOA varied pseudorandomly from trial to trial. Participants were instructed to fixate properly and to place their right hand in a central position. As soon as the visual target appeared, a saccade and/or a goal directed arm movement had to be made.

3 Experiment 1

The first experiment investigates whether the effects of spatiotemporal, visual-auditory relationship are the same for eye and arm movements.

3.1 Methods

Visual targets (19 cd/m^2) were presented at eccentricities of 15° or 25° to the left or to the right of a fixation point. Accessory auditory stimuli (76 dB) were presented straight ahead or 15° or 30° to the left or to the right; either 30 ms prior to the visual stimulus, simultaneously, or 60 or 120 ms after the onset of the visual stimulus.

3.2 Results

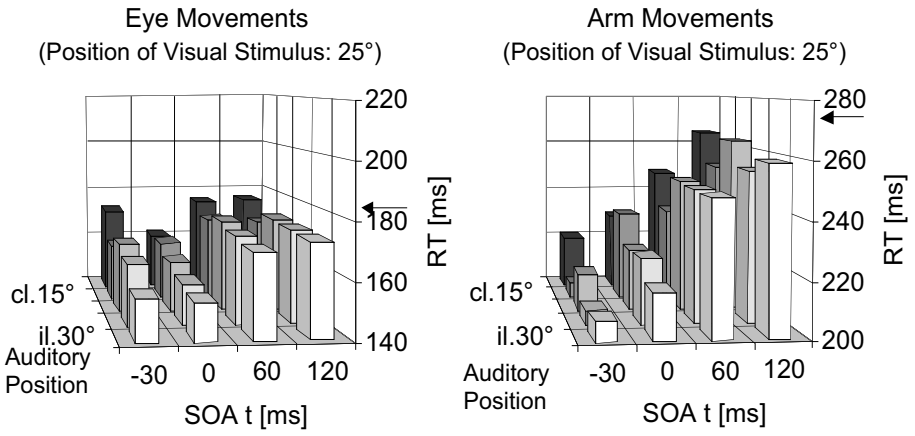


Fig. 1. Example of mean latencies to bimodal targets for different spatio-temporal stimulus conditions. The x-axis indicates the temporal arrangement (SOA), the z-axis refers to the spatial arrangement, i.e. relative position of the auditory stimulus with respect to the visual stimulus, il. = ipsilateral, cl. = contralateral. Mean unimodal reaction times are indicated by arrows.

However, the effect of the auditory signal was stronger for hand movement latencies than for eye movement latencies. This holds true for the latency reduction resulting from the presence of the auditory accessory and for the effects of spatiotemporal stimulus arrangement (Fig.1). Statistical analysis (ANOVA) revealed a significant interaction between type of movement (eye or arm) and effect of temporal resp. spatial stimulus arrangement. Moreover we observed a markedly higher number of directional errors in hand movements compared with eye movements - even if both movements had been performed simultaneously.

3.3 Discussion

The effect of SOA on reaction times may be resulting from unspecific warning effects elicited by the auditory signal. However, the dependency on spatial stimulus configuration shows that (a) general arousal or warning effect alone cannot account for the RT changes and that (b) visual and auditory information converges at some point of processing. Since these effects occur within eye and arm movements we may assume that visual-auditory integration follows the same rules in eye as in arm movement control.

On the other hand, our results suggest a stronger influence of the auditory stimulus on manual latencies. The stronger effect of SOA is easily explained by the fact that motor execution and muscle control are more complex in arm movements than in eye movements and thus may benefit more strongly from warning or arousal mechanisms. The assumption of a common control process is not violated. In contrast, the stronger effect of the spatial stimulus arrangement and the higher error rate in arm movements contradicts the hypothesis that eye and arm motor commands access the same multimodal representation of the environment.

4 Experiment 2

To corroborate the finding of Experiment 1 we varied the intensity of the accessory auditory stimulus. The hypothesis was that, due to the stronger dependence of arm movements on the auditory stimulus, the effect of intensity variations should be stronger for arm than for eye movements.

4.1 Methods

Possible positions for both stimuli were 25° to the left and to the right from the fixation point and, additionally, straight ahead for the auditory stimulus. Visual target and auditory signal were always presented simultaneously.
Visual stimulus intensity was 11 cd/m^2. Auditory intensities were determined individually for each subject in an intensity matching task. Three intensity levels were used: the intensity determined in the matching task and two additional intensities of 6 dB above and 6 dB below the determined intensity.

4.2 Results

As in the first experiment the spatial variation in the stimulus arrangement led to stronger effects on manual latencies compared with saccadic latencies. However, the latencies for both types of movement decreased with increasing auditory intensity in an almost identical manner. There was no evidence of a stronger effect of the auditory intensity on arm movements.

The comparison of eye movements with and without concomitant arm movements indicates an effect of the arm movements on visual-auditory integration in saccades.

A test designed to verify integration processes (horse race inequality test [6]) provides less evidence for multisensory integration in accompanied compared with unaccompanied saccades. Although a negative outcome of the test does not mean that no integration takes place, this finding would deserve further investigation.

4.3 Discussion

The lack of a stronger influence of auditory intensity on arm compared to eye movements contradicts the idea that, generally, the auditory stimulus has larger effects on arm movements compared to eye movements. Rather the result suggests that a higher spatial resolution for auditory stimuli in arm movement control evokes the differences found in Experiment 1. This raises the question what the origin of this higher resolution might be.

5 Experiment 3

A higher spatial resolution in arm movement control may be evoked by different representations of the visual-auditory environment for eye and arm movements or by the fact that arm movement latencies are approximately 100 ms longer than saccadic latencies. Although most of this latency difference is attributed to (peripheral) motor processes, it might provide additional processing time to improve auditory resolution. To investigate this hypothesis we varied SOAs over a wide range.

5.1 Methods

Visual and auditory stimuli were presented under free-field conditions either spatially coincident 25° to the left or right of the fixation point or in different hemispheres. Seven SOAs between –50ms (auditory first) and 250ms separated by 50ms were used. Auditory stimulus intensity was 65 dB SPL.

5.2 Results

For small SOAs the same spatiotemporal effects on eye and arm movement latencies were found as in Experiment 1. However, the spatial as well as the temporal effects decay for larger SOAs. In eye movements unaccompanied by arm movements the decay occurs at approximately 50 ms shorter SOAs compared to arm movements executed without eye movements. For conjoined eye and arm movements this difference is less clear.

Saccades are altered by the concomitant execution of an arm movement. Saccades are larger and in some subjects faster when accompanied by an arm movement compared with unaccompanied saccades. Amplitude, peak velocity and main

sequence (peak velocity with respect to saccadic amplitude) differ significantly between eye movements with and without concomitant arm movements [cf. 7].

5.2 Discussion

Similar effects of spatiotemporal stimulus arrangement on latencies of conjoined eye and arm movements are within expectation. The interdependence of the movements - reflected for example in changes in latencies and the main sequence data - may be the basis of this resemblance. However, the high degree of correspondence in *unaccompanied* eye and arm movements when corrected for the 50 ms difference in SOAs is remarkable. This may indicate, that both movements are based on the same representation of visual-auditory stimuli. Arm movement control processes might access the spatial representation approximately 50 ms later than systems controlling saccadic eye movements. During this additional processing time the spatial representation may have changed. E.g. the detection of auditory stimulus position which is based on the comparison of the input signals from the left and the right ear may be refined.

6 Conclusion

Our results corroborate the notion of a common control process in saccades and goal directed arm movements. Latency data suggest that both movements rely on the same visual-auditory representation which they access at different points in time.

References

1. Mushiake, H., Fujii, N., Tanji, J.: Visually Guided Saccade Versus Eye-Hand Reach: Contrasting Neuronal Activity in the Cortical Supplementary and Frontal Eye Fields. J. Neurophysiol. **75** (1996) 2187-2191
2. Werner, W., Dannenberg, S., Hoffmann, K.-P.: Arm-Movement-Related Neurons in the Primate Superior Colliculus and Underlying Reticular Formation: Comparison of Neural Activity with EMGs of Muscles of the Shoulder, Arm and Trunk During Reaching. Exp Brain Res **115** (1997) 191-205
3. Colonius, H., Arndt, P.A.: A Two-Stage Model For Visual-Auditory Interaction in Saccadic Latencies. Percept. Psychophys. **63** (2001) 126-147
4. Giray, M., Ulrich, R.: Motor Coactivation Revealed by Response Force in Divided and Focused Attention. J Exp Psychol HPP **19** (1993) 1278-1291
5. Meredith, M.A., Stein, B.E.: Visual, Auditory, and Somatosensory Convergence on Cells in Superior Colliculus Results in Multisensory Integration. J. Neurophysiol. **56** (1986) 640-662
6. Miller J.: Divided attention: evidence for coactivation with redundant signals. Cogn Psychol **14** (1982) 247-279
7. Snyder, L.R., Calton, J.L., Dickinson, A.R., Lawrence, B.M. : Eye-Hand Coordination : Saccades Are Faster When Accompanied by a Coordinated Arm Movement. J. Neurophysiol **87** (2002) 2279-2286

Probabilistic Integration of Cues From Multiple Cameras

J. Denzler[1], M. Zobel[1] and J. Triesch[2]

[1] Lehrstuhl für Mustererkennung
Universität Erlangen–Nürnberg
email: {denzler,zobel}@informatik.uni-erlangen.de
[2] Cognitive Science Department
University of California, San Diego
email: triesch@CogSci.ucsd.edu

Abstract Cue integration from multiple cameras is an important aspect for machine vision systems operating in complex, natural environments. One successful approach for self–organized cue integration is Democratic Integration. The hallmark of Democratic Integration is that different cues can autonomously determine whether and in how far they are useful for the current task, giving the system flexibilty to engage in different tasks and robustness in the face of sudden failures of cues. In this paper we embed Democratic Integration in a probabilistic framework and extend it hierachically in order to model *adaptive* cue integration for the general case of n calibrated cameras. Our experiments show that the method is capable of robust cue integration and adaptation during object tracking using three cameras placed arbitrarily in the scene.

1 Introduction

It is an unsolved problem in computer vision how sensor data selection and fusion should be done in the case that multiple cameras and multiple cues from each of the cameras are available. Such problems arise for example in surveillance tasks, where different sensors (e.g. infrared and daylight cameras) are placed at different positions in the environment and information from these sensors needs to be combined dependent on the environmental conditions (day/night, rain/sunshine, etc.). Also, the estimated position of the tracked object in the scene will have an influence on the contribution, each sensor can make. Of particular importance for real world applications in this respect is also, that individual sensors or cues may sometimes (unexpectedly) fail due to, e.g., limited view, occlusions, or hardware problems, or other reasons, and that the system must be robust with respect to such disturbances.

The main contribution of this work is a robust cue integration and adaptation mechanism for object tracking using multiple cameras. The basis of our approach is the Democratic Integration mechanism [3]. It is briefly summarized in the next section. Democratic Integration has originally been applied to fuse multiple cues arising from a single camera. We extend this approach towards hierarchically fusing cues originating from multiple calibrated cameras. Our goals are to demonstrate that cues from multiple cameras can be fused in a self-organized

manner, such that the contribution of each of the cameras is dependent on the estimated reliability of that camera, and that such a system is robust with respect to unexpected failure of individual cues or entire cameras.

2 Democratic Integration

The idea behind Democratic Integration is to integrate different perceptual cues in a self-organized manner [3]. Adaptation of the cues is driven by the agreement or compatibility between the different cues and sensors in the system. This idea was first studied in a face tracking system [3]. The system employed a stationary camera monitoring a room. Five simple cues analyzed the camera images. Each cue computes a 2-dim. *saliency map* registered to the camera image, in which high values inidicate a high confidence of the cue that there is a face at that location. The different cues are integrated or fused by computing a *result saliency map* which is a weighted average of the individual saliency maps. Importantly, the weights are time dependent and are constantly adpated in a self-organized fashion. To this end, an agreement or quality function is defined, that compares a cue's saliency map to the result saliency map. A cue whose saliency map is very similar to the result saliency map currently has a high quality. The important step now is to change the cue weights based on these qualities. A cue whose quality becomes very small, indicating disagreement of its saliency map to the result saliency map, will reduce its weight to no longer disrupt the overall system. Conversely, a cue that has recently been in very good agreement with the result will increase its weight. In addition, each cue can adapt internal parameters in order to better match its saliency map to the result saliency map. This allows the system to recalibrate cues and to use cues for a particular task that have no a priori information about the task. These cues are bootstrapped by other cues and simply adjust their internal parameters to match the result.

3 Probabilistic Fusion with Multiple Cameras

In Democratic Integration one of the key concepts is the result saliency map into which all different cues are fused to produce the final result for tracking with one camera. The main idea in our approach is, that for fusing the information gathered by multiple, calibrated cameras, the local and result saliency map is substituted by a probability distribution over a state space. Note, that it is quite intuitive to interpret the saliency map in 2–D — assuming proper normalization — as a distribution over a 2–D state space. In this special case the 2–D state consists of the position of the moving object on the image plane. In our approach we deal with the general case of an n–dimensional state space and observations that are made in several 2–D image planes.

The key idea of the hierarchical probabilistic approach can be summarized in the following informal way:

Probabilistic modeling of the state A particle filter framework is used to estimate the state of the object in 3–D (in the experiments the position, velocity, and acceleration of a moving object). This gives us a distribution over the state space represented by a particle set. A similar approach in the case of cue integration for a single camera has been proposed in [2].

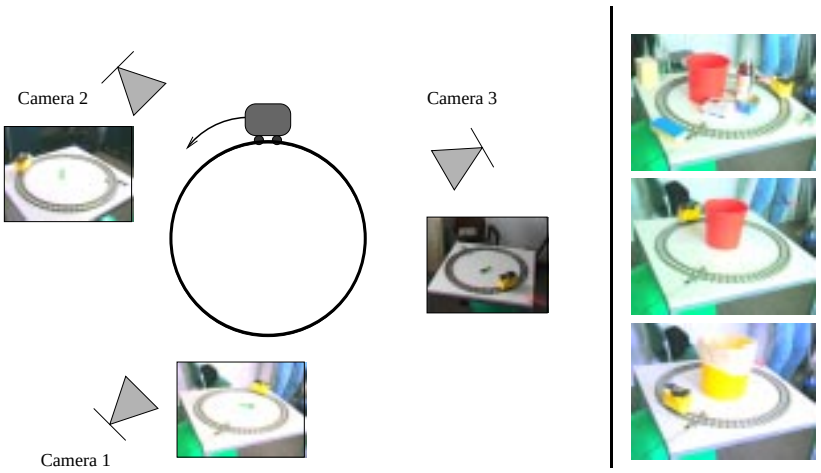


Figure 1. Left: Experimental setup. Position of the three cameras and the rail track. The images show the view on setup `basic`. Right: Images of setup `complex`, `bucket` and `yellow_bucket` (taken from camera 1).

Local state estimation For each sensor local state estimation is done using the original cue integration mechanism of Democratic Integration, i.e. a result saliency map is generated for each sensor from the different cues. This saliency map is used as likelihood function for evaluating the likelihood of each particle, that is drawn while applying the particle filter. In the case of calibrated cameras each particle, which might be interpreted as a kind of hypothesis for the 3–D state, is projected into the image plane and a score can be computed for each hypothesis by the likelihood function (for a detailed introduction on how particle filters are used the reader is referred to [1]). The weights of the different local cues as well as the other parameters of the cues are adjusted as described in [3] afterwards.

Global state estimation In an additional step a global state estimate is computed in a similar manner as it is done for each of the local state estimates. Each particle is projected onto the image planes of the different cameras. The global score of a particle is now computed as a weighted average of the local scores (already computed during the local state estimation). The weights, assigned to each camera, are updated in an additional Democratic Integration step. The main difference is, that now distributions represented as particle sets have to be compared, to figure out the agreement of the local estimates with the global ones. For comparison different metrics can be used to measure correspondence (agreement) between two distributions. One example is the Kulback–Leibler distance.

4 Experimental Setup and Results

During the experiments a moving toy train is tracked in 3-D using our proposed framework. 3-D estimation is conducted with a particle filter. The state (i.e. each

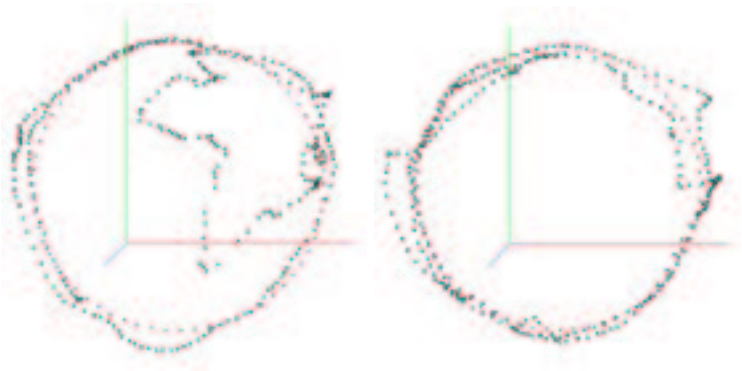


Figure 2. Estimated versus true motion path for setup `complex_occl`. Left: without sensor weight update. Right: with sensor weight update.

particle) consists of the 3-D position, velocity and acceleration of the object. For all experiments 2000 particle have been used.

In order to analyze our approach we choose the for the following basic experimental setup: the toy train is moving on a circular path in front of three cameras. Camera 1 and Camera 2 are SONY DFW–VL500 firewire cameras with a resolution of 320×240 at 25Hz. Camera 3 is a SONY digital camera with a resolution of 720×576 at 30Hz. The positions of the rail track and the three cameras are indicated in Figure 1. This setup is called `basic` in the following. In the beginning the cameras have been calibrated using Tsai's method [4].

Three different scenes are built up modifying the basic setup: a scene `complex` that contains a lot of different objects inside and outside the rail track to induce occlusions for one or the other camera and heterogeneous background. The scene `bucket` consists of a big red bucket in the center of the circular track, while in scene `yellow_bucket` a yellow bucket that has similar color as the moving toy train is used. Two more setups are constructed: `basic_occl` and `complex_occl`. In both cases the setups `basic` and `complex` are used, except for a sensor failure that was simulated by totally covering one of the cameras for a couple of seconds.

For each of the six setups a 10s sequence has been recorded for each of the three cameras simultaneously. The cameras have been manually synchronized only once at the beginning of the recording and in the end to subsample the 30Hz sequence of the third camera to match the 25Hz sequences of the first two cameras. The resolution of the images has been reduced to 80×60 for the first two cameras and to 75×60 for the third one. Additionally, the RGB images have been transformed to HSV color space.

To evaluate the quality of tracking for the different setups the circular rail track was reconstruced in 3–D using the calibration information of the cameras. As quality measure the mean euclidian distance between the estimated position of the toy train during tracking and the reconstruced circle in 3–D is used.

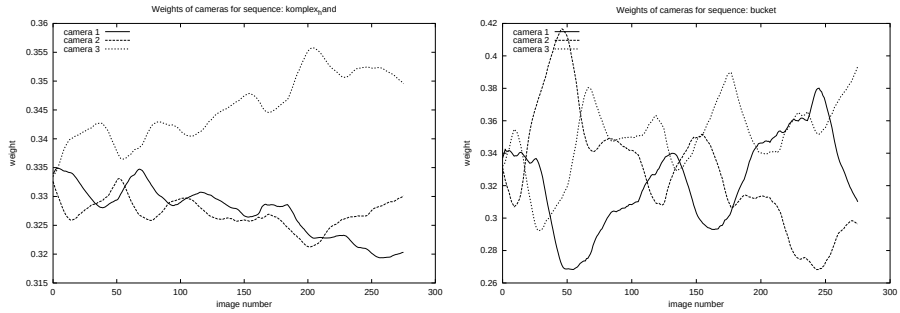


Figure 3. Cameras' weights for scenes `complex_occl` (left) and `bucket` (right)

For tracking the moving object, each camera uses the cues motion, prediction, contrast and color (for the computation and the parameters of these cues see [3]). Each experiment starts using only color and motion cue, i.e. the weights for color and motion cue are both set to 0.5. The other two cues are bootstrapped by the former ones.

In the experiments we tested different settings for the time constants τ_s (for sensor weight adaptation) and τ_c (for cue weight adaptation, see [3]). The time constants directly control how fast the influence of a sensor or a cue is changed. Since the different scenes differ in the demands on the adaptation, a compromise has been chosen between fast adaptation but not over–reacting on sensor noise or processing errors. Due to lack of space we only present results for $\tau_s = \tau_c = 10000$msec. Smaller values tend to improve the results for the sequences `complex` and `complex_occl` while at the same time the quality for `basic` and `bucket` is slighly reduced. For the setup `complex_occl` the advantage of the sensor weight adaptation can be best shown. Without sensor weight update tracking of the 3–D position breaks down during the simulated failure of sensor 1. With our proposed method (Figure 2, right) the system keeps track of the moving object with high accuracy. In Figure 3, left, the weights for cameras 1–3 are plotted over time. Evaluating the weights of the sensor over time, we can observe that the influence of each sensor is changed due to the visibility condition of the object (a periodic up and down of the weights can be observed). During failure of camera 1 the weight of this camera is decreased, as expected. A similar plot for scene `bucket` is shown in Figure 3, right, that again shows the periodic increase and decrease of the cameras' influence due to the visibility situation in the scene.

In Table 1 the estimation error is summarized for the different setups, Democratic Integration without and with sensor weight update as well as a result achieved if no cue and sensor adaptation is applied. In the latter case a non–adaptive particle filter approach is used to estimate the position in 3–D by probabilistic fusion of all three cameras.

	no weight update		weight update		no DI	
setup	mean	std. dev.	mean	std. dev.	mean	std. dev.
basic	24.6	14.4	22.3	13.0	39.1	23.7
bucket	22.7	13.3	26.6	16.6	50.7	34.2
yellow_bucket	46.7	32.5	38.8	28.3	130.4	73.4
complex	33.2	20.4	37.5	27.0	53.0	37.7
basic_occl	30.9	29.6	26.3	21.7	39.6	28.1
complex_occl	52.5	56.5	32.5	20.6	59.3	48.5
total	35.1		30.6		62.0	

Table 1. Mean euclidean error and standard deviation in the 3–D estimation of the moving toy train (in mm). Left column: without sensor weight update. Middle column: with sensor weight update. Right column: non–adaptive sensor data fusion using particle filters without adaptation of cues' or sensors' influence. The size of the toy train is approx. $110 \times 80 \times 90$mm at a distance of 1.5-2.0m from the cameras.

5 Conclusions

In this paper we have shown first, that the integration of cues from multiple cameras can be done very elegantly in a probabilistic framework using particle filters, and second, that adaptation in Democratic Integration can not only be performed locally in each sensor but also globally giving more influence to more reliable sensors at the current situation. The circumstances in our experiments (i.e. weak synchronisation of the cameras, different types of cameras, different and low resolution of the images) prove that our approach is robust and also capable for handling systematic differences in the reliablity of the sensors, as well as unexpected temporary failure of one or the other sensor[3]. The particle filter allows for handling multi–modal distributions over the state space, i.e. dealing with multiple hypotheses and objects in the scene.

Acknowledgment

The work was partially supported by the Bavaria California Technology Center under grant 2410-2001 and the German Science Foundation (DFG) under grant SFB603 TP B2.

References

1. A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, Berlin, 2001.
2. M. Spengler and B. Schiele. Towards robust multi–cue integration for visual tracking. In *ICVS 2001 Vancouver, Canada, 2001*, pages 93–106. Springer, 2001. Lecture Notes in Computer Science.
3. J. Triesch and C. von der Malsburg. Democratic integration: Self-organized integration of adaptive cues. *Neural Computation*, 13(9):2049–2074, 2001.
4. R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal of Robotics and Automation*, Ra-3(3):323–344, August 1987.

[3] The reader is referred to http://www5.informatik.uni-erlangen.de/~di for image sequences and results of the processed scenes

Sensor Fusion for Vision and Sonar Based People Tracking on a Mobile Service Robot

T. Wilhelm, H.-J. Böhme and H.-M. Gross

Ilmenau Technical University, Department of Neuroinformatics, P.O.Box 100565, 98684 Ilmenau
{wilhelm, hans, homi}@informatik.tu-ilmenau.de

Abstract. Service robots intended to interact with people must be able to localize and continuously track their users. A method is described which integrates information from visual and sonar based tracking pathways while updating hypotheses about the position of the robot's human user. Each tracking method uses information from the other to generate a more robust measure of the user's position, and thus a more robust behavior generation is achieved.

1 Introduction

A service robot, which is designed to serve people in special domains or to help them in their everyday life, must be able to localize and continuously track its users. If the user breaks the interaction off, there is no need for the robot to continue to produce any outputs. Lacking these capability would result in a robot, which is trying to contact arbitrary things or which is proceeding to offer its services even when the user already left the operation area. The authors consider the knowledge about the position of the user as fundamental for a smart appearance of any service robot. On the other side, the price determines the economical success of any service robot application, so it seems favorable to use cheap hardware whenever possible, which has consequences on the complexity of any people tracking algorithm.

Our experiments were carried out in a home store, where our service robot is to operate as a mobile shopping assistant, guiding customers to desired products in the store [1]. A major problem concerning people tracking in this environment are the varying illumination conditions from natural to artificial lighting, which imply a multimodal approach to the problem, not only relying on visual cues.

2 Tracking

Tracking of users can be realized by using different sensor systems. The distance to an object can be measured by means of sonar or laser data, and there are methods that extract hypotheses about the position of people in the robot's surroundings from laser data [6]. In contrast to laser scanners, the resolution and accuracy of sonar sensors give only a vague hint about the nature of the object,

and it seems that these methods can not be assigned to cheap sensor systems such as sonars. Moreover, the used features are not very person-specific and could detect other objects as potential users as well. Cameras can be considered as cheap sensors compared to laser scanners, and visual data can be used to solve ambiguous situations and to discriminate people from arbitrary objects. Thus the proposed tracking method consists of a sonar and a vision based tracking module.

2.1 Sonar Based Tracking

The task of the sonar based tracking is to always keep contact to the user by moving the robot according to its mode of operation and the position of the user. Our experiments were carried out on a B21 mobile robot (RWI IS Robotics) equipped with two layers of sonar sensors with 24 sonars respectively. The raw sensor data is noisy and depends on the orientation and the material of the objects around the robot. Therefore the raw data is preprocessed as follows:

1. replacement of invalid measurements: distances larger than $22,5m$ are considered as invalid and are replaced by the previous measurements
2. local spatial low pass filtering of adjacent measurements
3. temporal low pass filtering of successive measurements
4. calculation of a weighting factor in each direction which is inversely proportional to the measured distance $W_{Sonar}^{(c)} = 1 - d_{sonar}^{(c)}/d_{max}$, where $d_{sonar}^{(c)}$ is the preprocessed sonar measurement at position c in the scan and d_{max} is the maximum distance $(1,5m)$; for distances larger than d_{max} the weight is set to 0

The position of the maximum in the resulting weighting vector corresponds to the nearest object (see Figure 2e) and is used to generate an appropriate behavior, depending on the robots mode of operation:

1. *communication:* orient the touch interface mounted on top of the robot to the position of the maximum, thus allowing the user to make inputs
2. *guide user:* keep the distance to the user small and stay in front of him, while driving towards a goal position in the market
3. *follow user:* keep distance to user small and try to stay behind him

The advantages of the sonar based tracking are its low computational costs and thus its ability to continuously track the user and align the robot appropriately. It generates an adequate behavior as long as the nearest object is really the user, otherwise the robot reacts to any object in its surroundings and tries to interact with it. This drawback can be encountered by integrating information from a vision based tracking module, which is able to distinguish people from any other objects in the area.

2.2 Vision Based Tracking

The basis of the vision based tracking procedure is the condensation algorithm [4]. The task of calculating the probability of the existence of a person for every pixel and tracking the resulting density function is solved by an approximation of the density function by a relatively small number of samples. The condensation algorithm operates on the panorama images from an omnidirectional color camera and uses different feature extraction methods to calculate hypotheses about humans faces and the upper part of the human body. Compared to a panoramic image with 720×106 pixels calculating the feature extraction only for 200 samples yields a reduction to merely 0.262%.

Skin Color A widely used method for finding faces in images is skin color classification. Here the dichromatic r-g-color space ($r = R/(R+G+B)$, $g = G/(R+G+B)$) is used, which is widely independent from variations in luminance. The color model consists of a look up table with manually classified skin color pixels in the r-g-color space [3]. To prevent the color model from getting holey because of insufficient training data, there is a small Gaussian placed around each skin color pixel. The skin color model is depicted in Figure 1. The color detection can be calculated very fast but it is highly dependent on illumination color and variations in hue and often fails in back light situations.

Head-Shoulder-Contour The second method uses a contour model which describes the mean head-shoulder-contour of a person [2]. The model Λ was derived from a number of images containing frontally aligned persons. On the mean gray level image, the local orientations were calculated with a structure tensor [5]. The same tensor is used during head-shoulder-contour detection to calculate the gray value orientation in a local surrounding around each sample, and the template matching is carried out for every sample according to equation 1, where o is the orientation in the image and λ is the orientation in the contour model. Figure 1 depicts the head-shoulder-contour model Λ of size 20×20.

$$W_{hsc}(x,y) = \frac{\sum_{i=0}^{I-1} \sum_{j=0}^{J-1} \frac{1}{2} \left[\cos(2|\lambda_{i,j} - o(x-i, y-j)|) + 1\right]}{card(supp(\Lambda))} \tag{1}$$

The head-shoulder-contour is computational more expensive and not as person specific as the skin color detection, but it yields good results in back light situations, where any other gray value or color based face detector fails.

Combination of the Vision Based Cues Although both cues are person-specific, it can happen that they do not detect a user or give false alarms. Therefore both cues are combined by a fuzzy min-max-operator ($minmax(a,b) = \gamma \min(a,b) + (1-\gamma)\max(a,b)$), which can be configured between a pessimistic and an optimistic fusion. Pessimistic (min, $\gamma = 1$) means that an user which was not detected by at least one cue is not accounted for at all, while using the

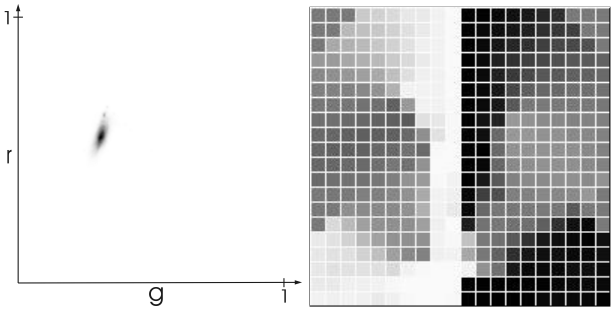


Fig. 1. Models used for vision based tracking. *Left*: Color model in the dichromatic r-g-color space. *Right*: Head-shoulder-contour model, local orientations are represented by gray levels, where white and black pixels code horizontal, and gray pixels code vertical edges respectively

max ($\gamma = 0$) fusion results in a behavior, where all false positive matches from one cue are considered valid. See Figure 2 for results of the single cues and their combination.

3 Sensor Fusion

As mentioned before, vision based tracking shall now be used to prevent the sonar based tracking from interacting with arbitrary non-human objects. On the other hand, the vision based tracking can benefit from the sonar based method by using it as third cue for calculating the sample weighting.

Support of Vision Based Tracking by Sonar Data Since the sonar scan as well as the image constitute an 360° description of the robots surroundings, it is possible to assign a scan measurement at position c in the scan to each position $\boldsymbol{x}$ in the image. This way, the sonar vector can be used to modulate the sample weighting in the condensation algorithm, equation 2 and 3. Thus only those samples get a high weight, that are supported by the vision based cues and, at the same time, lie in a direction with a short distance measured from the sonar sensors. Samples that are only supported either by the vision or the sonar based tracking eventually die out (Figure 3).

$$W^{(i)}_{Sample}(\boldsymbol{x}) = minmax\left(W^{(i)}_{skincolor}(\boldsymbol{x}), W^{(i)}_{hsc}(\boldsymbol{x})\right) W_{sonar}(c) \tag{2}$$

$$P^{(i)}_{Sample}(\boldsymbol{x}) = \frac{W^{(i)}_{Sample}(\boldsymbol{x})}{\sum_i W^{(i)}_{Sample}(\boldsymbol{x})} \tag{3}$$

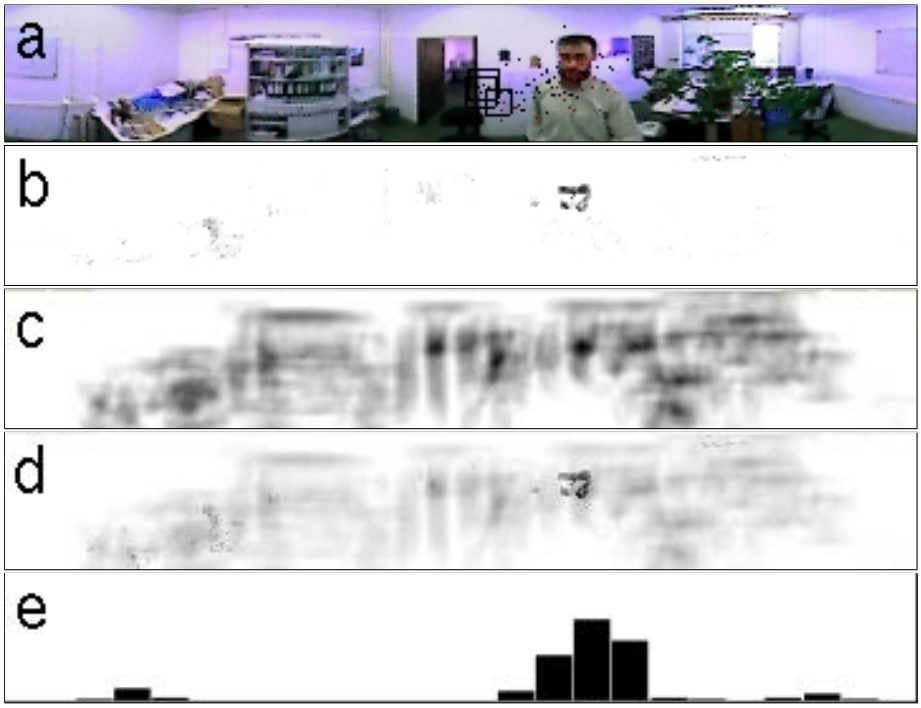


Fig. 2. Results of the single vision based tracking modules and the sonar based tracking: a) original panoramic image; b) skin color classification; c) head-shoulder-contour detection; d) $MinMax$-fusion ($\gamma = 0.7$), note that at the position of the users head, both cues give the largest contribution; e) weighting factors W_{hsc} calculated from the sonar scan

Support of Sonar Tracking by Vision Based Data Since only the sonar based tracking is responsible for behavior generation, the case where vision based data supports sonar tracking is more important. The camera image is divided into columns corresponding to the single sonar measurements. In every column c, the sum of the sample weights is calculated, resulting in a vector with high values on those positions where most likely the user is. For behavior generation, the positions of the maxima in the sonar and vision based scan are compared. If they are aligned, the motor commands are executed, otherwise all actions are suppressed. Thus, other people can approach, without the robot turning away from its current user.

4 Summary

The paper presents the integration of a sonar and a vision based user tracking pathway into a robust tracking procedure, which was applied successfully on a mobile service robot in a home store.

Fig. 3. Comparison of pure vision based tracking (*left*) and vision based tracking with sonar support (*right*). Every 10th image in the sequence is shown; the user moves around the robot (sometimes the robot is turning to the user based on sonar tracking). While at the left many samples get stuck on other objects, the tracking with sonar support does not loose the user

5 Outlook

In our current work, we investigate possibilities of automatic camera color calibration to get the skin color classification independent from variations in illumination color. In addition to that, we analyse the performance of other feature extraction and face detection methods, such as cascade correlation neural networks for the vision based tracking pathway. Furthermore, a robotic face with two cameras was designed, which is always oriented towards the currently tracked person. High resolution images from these frontally alligned cameras can be used to recognize a user who was lost from the omnidirectional view during tracking.

References

1. Boehme, H.-J., Wilhelm, T., Key, J., Schroeter, Ch., Hempel, T., and Gross, H.-M. An Approach to Multimodal Human-Machine Interaktion for Intelligent Service Robots. In *EUROBOT'01 - the fourth Euromicro Workshop on Advanced Mobile Robots*, volume 86 of *Lund University Cognitive Studies*, pages 17–24. Lund University, 2001.
2. Corradini, A., Boehme, H.-J., and Gross, H.-M. A Hybrid Stochastic-Connectionist Approach to Gesture Recognition. *International Journal on Artificial Intelligence Tools*, 2000(9):177–204, 2000.
3. Feyrer, S. *Detektion, Lokalisierung und Verfolgung von Personen mit einem mobilen Serviceroboter*. PhD thesis, Eberhard-Karls-Universität Tübingen, 2000.
4. Isard, M. and Blake, A. CONDENSATION – conditional density propagation for visual tracking. *International Journal on Computer Vision*, 29(1):5–28, 1998.
5. B. Jähne. *Digitale Bildverarbeitung*. Springer-Verlag, Berlin Heidelberg, 3. Auflage, 1993.
6. Schulz, D. and Burgard, W. Probabilistic state estimation of dynamic objects with a moving mobile robot. *Robotics and Autonomous Systems*, 34:107–115, 2001.

A Stochastic Model of Multimodal Integration in Saccadic Responses

Hans Colonius[1] and Adele Diederich[2]

[1] Oldenburg University, P.O. Box 2503, D-26111 Oldenburg, Germany,
hans.colonius@uni-oldenburg.de,
http://www.psychologie.uni-oldenburg.de/hans.colonius/index.html
[2] International University Bremen, School of Humanities and Social Sciences,
Campus Ring 1, 28759 Bremen
a.diederich@iu-bremen.de,
http://www.iu-bremen.de/hss/adiederich/

Abstract. The time-window-of-integration model is a quantitative framework for describing crossmodal effects in saccadic response time. It distinguishes a first stage of parallel peripheral processing followed by a second stage of multimodal integration. The occurrence of crossmodal effects (facilitation/inhibition) hinges upon the peripheral processes terminating within a temporal window of integration. The window mechanism is determined by unimodal stimulus properties like intensity, while the size of the effect is modulated by crossmodal stimulus properties like spatial configuration.

1 Multimodal Integration in Saccadic Responses

Saccades are fast, voluntary movements of the eyes to align the high-resolution fovea with objects and events of interest. In a natural environment saccades are part of a rapid goal-directed orienting response system to stimuli occurring in the periphery. Stimuli are usually multimodal: in addition to visual and auditory inputs, vestibular and somatosensory afferents have access to the saccade-generating mechanism. Thus, the oculomotor system has become a prominent site for the analysis of crossmodal integration.

For example, it has been found that saccadic reaction time to visual targets (the time between the onset of the visual stimulus and the onset of the saccadic eye movement) tends to be faster when auditory stimuli are presented in close temporal or spatial proximity (see [1], [2], [3], [4]). Similar response enhancement effects for saccades have been observed for combining visual and somatosensory stimuli (cf. [5] for monkeys; [6] for humans).

These behavioral studies are in line with neurophysiological evidence for multisensory integration in the deep layers of the superior colliculus (DLSC) (see [7], [8]). Multisensory neurons in DLSC of anaesthetized cats ([9]) and monkeys ([10]) showed an enhanced response to particular combinations of visual, auditory, and tactile stimuli paralleling the spatial-temporal rules in the behavioral

studies. Similar results for recordings from the awake behaving monkey have recently been obtained ([11], [12]).

Here we present a stochastic model that establishes a formal framework in which rules of crossmodal integration can be stated. Within that framework one can specify how the integration mechanism depends on the uni- and multimodal stimulus parameters and on specifics of the experimental paradigm. It allows to make qualitative and quantitative predictions and should thereby ultimately provide a link between the neural and the behavioral level of investigation.

2 The Time-Window-of-Integration (TWIN) Model

Since stimulation from different modalities like vision and touch cannot interact (e.g., on the retina), the model claims the existence of a *first stage* of parallel independent modality-specific activations in the afferent pathways. It refers to a very early stage of processing where detection of the stimuli, but possibly no "higher" processes like localization and identification, take place. This does not preclude the possibility of interaction between modality-specific pathways, nor between modality-specific and crossmodal areas, at a later stage of processing. In fact, there is increasing evidence that crossmodal processing does not take place entirely in feedforward convergent pathways but that it can also modulate early cortical unisensory processing ([13]). Thus the entire processing time must consist of at least two stages arranged in series. The *second stage* comprises neural integration of the input and preparation of the ocular motor response. True interaction, however, resulting in facilitation or inhibition of the response is supposed to occur only if the peripheral processes of the first stage all terminate within a given temporal "window of integration".

Even under invariant experimental conditions, saccadic responses typically vary from one trial to the next due to an inherent variability of the underlying neural processes in both ascending and descending pathways. This is taken into account by assuming the duration of each of the stages to be a random variable.

2.1 Distribution-Free Model Properties and Predictions

According to the model, observed reaction time in the multimodal condition can be written as a sum of two nonnegative random variables with finite first and second moments:

$$RT_{multimodal} \stackrel{d}{=} W_1 + W_2, \tag{1}$$

where W_1 and W_2 refer to first and second stage processing time, respectively[3]. Let I denote the event that crossmodal interaction occurs, having probability P[I]. Thus the saccadic response time (SRT) distribution is a binary mixture of two distributions defined by conditioning on event I:

[3] $\stackrel{d}{=}$ stands for "equal-in-distribution".

$$\begin{aligned} \mathrm{P}[RT_{multimodal} \leq t] &= \mathrm{P}[W_1 + W_2 \leq t] \\ &= \mathrm{P}[I]\mathrm{P}[W_1 + W_2 \leq t|I] + (1 - \mathrm{P}[I])\mathrm{P}[W_1 + W_2 \leq t|\text{not-}I]. \end{aligned} \tag{2}$$

While neither P[I] nor the two conditional distributions in Eq.(2) can be estimated directly from the data, mixture distributions have several distinctive properties that lead to empirically testable predictions even if no specific distribution assumptions are introduced (see [14]).

For the expected SRT in the multimodal condition then follows:

$$\begin{aligned} \mathrm{E}[RT_{multimodal}] &= \mathrm{E}[W_1] + \mathrm{E}[W_2] \\ &= \mathrm{E}[W_1] + \mathrm{P}[I]\mathrm{E}[W_2|I] + (1 - \mathrm{P}[I])\mathrm{E}[W_2|\text{not-}I] \\ &= \mathrm{E}[W_1] + \mathrm{E}[W_2|\text{not-}I] - \mathrm{P}[I](\mathrm{E}[W_2|\text{not-}I] - \mathrm{E}[W_2|I]), \end{aligned}$$

where E[$W_2|I$] and E[$W_2|$not-I] denote the expected second stage processing time conditioned on interaction occurring (I) or not occurring (not-I), respectively. Putting $\Delta \equiv \mathrm{E}[W_2|\text{not-}I] - \mathrm{E}[W_2|I]$, this becomes

$$\mathrm{E}[RT_{multimodal}] = \mathrm{E}[W_1] + \mathrm{E}[W_2|\text{ not-}I] - \mathrm{P}[I]\ \Delta. \tag{3}$$

The product P[I] Δ is a measure of the expected crossmodal interaction in saccadic RT in the second stage, with positive Δ values corresponding to facilitation, negative ones to inhibition.

In the unimodal conditions, no interaction is possible. Thus,

$$\mathrm{E}[RT_{unimodal}] = \mathrm{E}[W_1] + \mathrm{E}[W_2|\text{not-}I], \tag{4}$$

and the amount of crossmodal interaction is

$$\mathrm{E}[RT_{unimodal}] - \mathrm{E}[RT_{multimodal}] = \mathrm{P}[I]\ \Delta.$$

Several empirically testable predictions can now be formulated. First, the amount of crossmodal interaction should depend on the stimulus onset asynchrony (SOA) between the stimuli. For example, a stimulus with faster peripheral processing has to be delayed in such a way that the arrival times of both stimuli have a higher probability of falling into the window of integration. Indeed, the effect of crossmodal interaction tends to be most prominent when there is some characteristic temporal asynchrony between the stimuli ([1],[2]). Second, the probability of interaction, P[I], should depend on unimodal features that affect the speed of processing in the first stage, like stimulus intensity or eccentricity. For example, if a stimulus from one modality is very strong compared to the other stimulus' intensity, the chances that both peripheral processes terminate within the time window are small (assuming simultaneous stimulus presentations). The resulting low value of P[I] is in line with the empirical observation that a very strong target signal will effectively suppress any interaction with other modalities. The principle of "inverse effectiveness", according to which crossmodal facilitation is strongest when stimulus strengths are weak or close to

threshold level ([9]), can be accommodated in the model by adjusting the width of the time window: for low-level stimuli the window should become larger so as to increase the likelihood of crossmodal integration. Third, the amount of crossmodal interaction (Δ) and its direction (facilitation or inhibition) occurring in the second stage depend on crossmodal features of the stimulus set, in particular spatial disparity and laterality[4]. On the other hand, crossmodal features have no influence on first stage processing time since the modalities are yet being processed in separate pathways.

More specific predictions concerning the effects of varying stimulus intensity are implied by the following rules governing the window-of-integration mechanism. When the task is to orient toward the target modality stimulus ignoring stimuli from other modalities (*focused attention*), the first stage duration is determined by the target peripheral process, but crossmodal integration is occurring only if the non-target stimulus wins the race in the first stage, i.e., the window of integration is opened only by activity triggered by the non-target stimulus. Increasing the intensity of the target stimulus will thus increase its chances to win the race decreasing the probability that the window of integration opens, so that less crossmodal interaction should occur. This prediction is in line with the observation that a very strong target signal will suppress any interaction with other modalities. Increasing the intensity of the non-target stimulus, however, leads to the opposite prediction: the non-target stimulus will have a better chance to win the race and to open the window of integration, hence predicting more crossmodal interaction on average. On the other hand, when the task is to orient toward the first stimulus detected no matter of which modality (*redundant target*), the first stage duration is determined by the winner's peripheral processing time, and the window of integration is opened by whichever stimulus wins the race. Here, the effect of stimulus intensity depends on additional assumptions not outlined here.

2.2 TWIN Model with Exponential First Stage Distributions

The peripheral processes in the first stage are assumed to have stochastically independent exponentially distributed durations. The exponential assumption is motivated by mathematical simplicity and, together with a Gaussian distribution assumption for second stage processing time, results in an Ex-Gaussian distribution from that has been demonstrated to be a reasonably adequate description for many empirically observed reaction time data (cf. [15]). To illustrate the derivation for the expected SRT, consider a focused attention experiment with a visual target and an auditory non-target stimulus. The first stage duration is determined by the target peripheral process of random duration V, say, yielding $\mathrm{E}[W_1] = \mathrm{E}[V] = 1/\lambda_V$ (λ_V denotes the intensity parameter of the exponential distribution of V). From the assumptions stated in the last section,

$$I = \{A + \tau < V < A + \tau + \omega\} \tag{5}$$

[4] Laterality here means whether or not all stimuli appear in the same hemisphere.

where A is the peripheral auditory latency and τ and ω denote SOA and window width, resp. Straightforward calculation yields

$$\mathrm{P}[I] = \frac{\lambda_A}{\lambda_A + \lambda_V} \left\{ \exp[-\lambda_V \tau] - \exp[-\lambda_V (\tau + \omega)] \right\}, \tag{6}$$

where λ_A refers to the auditory intensity parameter. It is obvious from Eq. (6) that the probability of interaction increases both with λ_A and the window width ω, as it should. Expected saccadic reaction time then is

$$\mathrm{E}[RT_{multimodal}] = 1/\lambda_V + \mu - \frac{\Delta\, \lambda_A}{\lambda_A + \lambda_V} \left\{ \exp[-\lambda_V \tau] - \exp[-\lambda_V (\tau + \omega)] \right\}$$

where $\mu = \mathrm{E}[W_2|\text{not-}I]$, the mean duration of the second stage when no interaction occurs.

The choice of the second stage distribution is irrelevant as long as only mean latencies are considered. For predictions of the entire saccade latency distribution it should be noted, however, that due to conditioning on the event of interaction I the two stage durations W_1 and W_2 are not stochastically independent. For the model version considered in this section, it can be shown that they are negatively dependent if Δ is positive: in any given trial, whenever the visual peripheral process ($V \equiv W_1$) is relatively slow, the auditory peripheral process has a better chance of winning the race and opening the integration window, thus increasing the likelihood of facilitation in the second stage, and vice versa.

3 Conclusion

The TWIN model has recently been shown to give an excellent description of crossmodal effects on SRT in visual-auditory and visual-tactile focused attention experiments ([1],[6]). Note, however, that it is not meant to mirror multisensory processes at the level of an individual neuron. There are many different types of multisensory convergence occurring in individual neurons (see [16]), and some of their activities are consistent with the TWIN assumptions while others are not. Note also that the two-stage assumption does not preclude the possibility of interaction between modality-specific pathways, nor between modality-specific and crossmodal areas, at a later stage. In future work, the second stage mechanisms should be specified in more detail, in particular with respect to the spatial stimulus configuration effects. There is a large data base on receptive field properties of multisensory neurons available now (cf. [17]), and connecting these with behavioral data via an appropriate elaboration of the TWIN model should be a challenging task.

Notes and Comments. This work was supported by DFG grants Di 506/7-2 and SFB 517/C3 (Neurokognition).

References

1. Colonius, H., Arndt, P.: A two-stage model for visual-auditory interaction in saccadic latencies. Perc. Psychophys. **63** (2001) 126–147
2. Frens, M.A., Van Opstal, A.J., Van der Willigen, R.F.: Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. Perc. Psychophys. **57** (1995) 802–816
3. Harrington, L.K., Peck, C.K.: Spatial disparity affects visual-auditory interactions in human sensorimotor processing. Exp. Brain Res. **122** (1998) 247–252
4. Hughes, H.C., Nelson, M.D., Aronchick, D.M.: Spatial characteristic of visual-auditory summation in human saccades. Vis. Res. **38** (1998) 3955–3963
5. Groh, J.M., Sparks, D.L.: Saccades to somotosensory targets. I. Behavioral characteristics. J. Neurophys. **75** (1996) 412–427
6. Diederich, A., Colonius, H., Bockhorst, D., Tabeling, S.: Visual-tactile interaction in saccade generation. Exp. Brain Res., to appear
7. Munoz, D.P., Wurtz, R.H.: Saccade-related activity in monkey superior colliculus. I. Characteristics of burst and buildup cells. J. Neurophysiol. **73** (1995a) 2313–2333
8. Munoz, D.P., Wurtz, R.H.: Saccade-related activity in monkey superior colliculus. II. Spread of activity during saccades. J. Neurophysiol. **73** (1995b) 2334–2348
9. Meredith, M.A., Stein, B.E.: Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. J. Neurophysiol. **56** (1986) 640–662
10. Wallace, M.T., Wilkinson, L.K., Stein, B.E.: Representation and integration of multiple sensory inputs in primate superior colliculus. J. Neurophysiol. **76** (1996) 1246–1266
11. Bell, A.H., Corneil, B.D., Meredith, M.A., Munoz, D.P.: The influence of stimulus properties on multisensory processing in the awake primate superior colliculus. Can. J. Exp. Psych. **55:2** (2002) 123–132
12. Frens, M.A., Van Opstal, A.J.: Visual-auditory interactions modulate saccade-related activity in monkey superior colliculus. Brain Res. Bull. **46** (1998) 211–224
13. Driver, J., Spence, C.: Multisensory perception: Beyond modularity and convergence. Curr. Biol. **10** (2000) 731–735
14. Yantis, S., Meyer, D.E., Smith, J.E.K.: Analysis of multinomial mixture distributions: New tests for stochastic models of cognition and action. Psych. Bull. **110** (1991) 350–374
15. Luce, R.D.: Response Times: Their Role in Inferring Elementary Mental Organization. Oxford University Press, New York (1986)
16. Meredith, M.A.: On the neural basis for multisensory convergence: a brief overview. Cog. Brain Res. **14** (2002) 31–40
17. Kadunce, D.C., Vaughan, J.W., Wallace, M.T., Stein, B.E. The influence of visual and auditory receptive field organization on multisensory integration in the superior colliculus. Exp. Brain Res. **139** (2001) 303–310

Multimodal representations for human 3D object recognition

Ingo Rentschler[1], Martin Jüttner [2], Erol Osman[1], Alexander Müller[1], Terry Caelli[3]

[1] Institute of Medical Psychology, University of Munich, Munich, Germany
ingo@imp.med.uni-muenchen.de; erol@lrz.uni-muenchen.de;
alexkarl.mueller@arcormail.de
[2] Neuroscience Research Institute, Aston University, Birmingham, U.K.
m.juttner@aston.ac.uk
[3] Department of Computer Science, University of Alberta, Edmonton/Alberta, Canada
tcaelli@ualberta.edu

Abstract. School-children were provided with controlled prior knowledge about otherwise unfamiliar 3D objects, the effect of which on visual category learning and generalisation for 3D objects from images was examined. There occurred a developmental double-dissociation in that at age 8-9 years visual prior knowledge reinforced visual category learning more than haptic prior knowledge did, whereas at age 13–14 years and beyond haptic prior knowledge was much more effective. Generalisation performance revealed that object representations were view-based for the youngest children but multimodal for adolescents and adults. It is suggested that haptic prior knowledge reinforces visual object recognition by facilitating the solution of the correspondence problem for matching input data to internalised 3D object representations.

1 Introduction

For retrieving the spatial structure of 3D objects biological vision systems need to rely on additional information not given in the static retinal image. Thus it is generally assumed that image data are referenced to object representations stored in memory and current models of human object memory differ in their degree of view-point independence and view-point dependence. View-point independence is claimed by the recognition-by- (3D) components model, a non-algorithmic account of how the visual input is related to the non-accidental characterisation of object parts and their relations [1]. On the other hand, the multiple-views hypothesis holds that a set of object views is stored in memory and the object is recognised by relating the input view to the nearest view stored in memory [2]. These approaches focus on the variation of observer performance with varying views and ignore the fact that, due to object redundancies, view-point dependence of performance cannot be equated with that of representation [3]. They also ignore the need for a matching process between mental representations and input images [4].

We aimed at contributing to these issues by using an experimental paradigm characterised by several features. First, the effects of input information and object

representation were separated by providing humans with a controlled amount of prior knowledge about otherwise unfamiliar objects and then submitting them to a fixed procedure of visual category learning and generalisation [5]. Second, we explored how object information relevant to visual object recognition can be created in human memory with haptic and with visual input. Third, we investigated in a developmental context how humans learn with a teacher (*supervised learning*) to achieve the categorisation of 3D objects from 2D views [6]. Fourth, by using objects carefully constructed to have view-dependent symmetries and ambiguities we have excluded the possibility that recognition can be based solely on image matching (as in cross-correlation between view-dependent objects in memory and images).

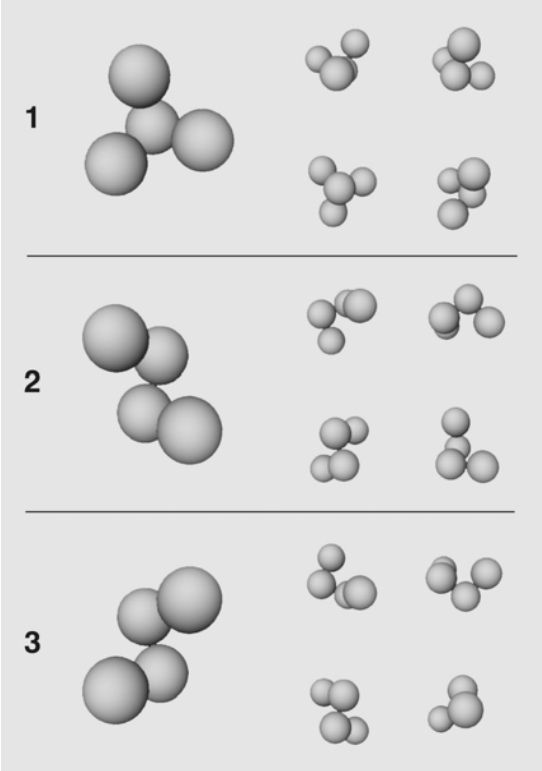


Fig. 1. The 3D objects used (left) and examples from the learning set of 2D views used for category learning (right).

2 Method

The three learning objects each consisted of three spheres forming an isosceles triangle and a fourth sphere placed upright above the centre of one of the base

spheres. Objects 2 and 3 were mirror symmetric to each other (Fig. 1). Real object models were constructed of polystyrene balls, measuring 6 cm in diameter. Virtual object models were constructed and displayed on a O2 workstation (Silicon Graphics Inc., USA). The learning set of 22 object views sampled the viewing sphere in 60° steps, the test set of 64 novel views in 30° steps. On the computer screen the objects subtended 1.5° of visual angle at a viewing distance of 1 m.

Learning units consisted of a learning phase and a test phase. During learning phases subjects saw the learning set three times in random order paired with class labels. Test phases consisted of classifying the learning set once (for details see ref. 6). For acquiring prior object knowledge subjects were assigned to either a control, a visual, or a haptic group. Subjects of control groups received no prior knowledge. Subjects of visual groups rotated the virtual object models on the computer screen via the mouse. The blindfolded subjects of the *haptic* groups were encouraged to freely manipulate the real object models.

Three groups with a total of 45 school-children and a control group of 15 adults participated. Thirty children were in elementary school, grade 3 (8-9 years) and grade 4 (9-10 years), fifteen in high-school, grade 8 (13-14 years). The fifteen adults were aged 20-45 years. The age groups were equally distributed over the three conditions of prior knowledge.

3 Results

Fig. 2 shows that the type of prior had a distinct impact on visual category learning. At age 13-14 years children with visual prior knowledge were not significantly better than children at age 8-9 years. In the same period there was little change in learning performance for the control groups as well. By contrast, the children with haptic prior knowledge were at age 13-14 years significantly better in learning performance than at age 9-10 years and much better than at age 8-9 years. MANOVA tests on the maximum performance and the average performance yielded significant effects of the factors age and condition, as well as a significant interaction between these factors.

The generalisation to novel object views is shown in Fig. 3. Performance is measured in terms of signal detection d' separately for the "non-symmetric" object 1 and for the "symmetric" objects 2 and 3. In the latter case, d'-values are mean values for objects 2 and 3. The children of the control groups show independently of age no generalisation abilities at all. The children in grades 4 and 8 show a similar improvement in generalisation due to visual and haptic prior knowledge. The generalisation performance of all age groups of school-children is clearly worse for the symmetric objects as compared to the non-symmetric object. This difference in generalisation performance is to some extent overcome by the adults.

The complete lack of generalisation by the children of the control groups suggests that they simply learned to associate the views of the learning set to object classes. By contrast, the convergence of "visual" and "haptic" generalisation performance for symmetric objects at grades 4 and 8, as well as for non-symmetric objects at grade 8,

indicates the development of visuo-haptic object representations with balanced information input from the two modalities. The fact that childrens' generalisation performance for the mirror-image objects was generally worse than that for the non-symmetric object suggests that learning relational object representations poses greater demands on cognitive development than learning non-relational representations.

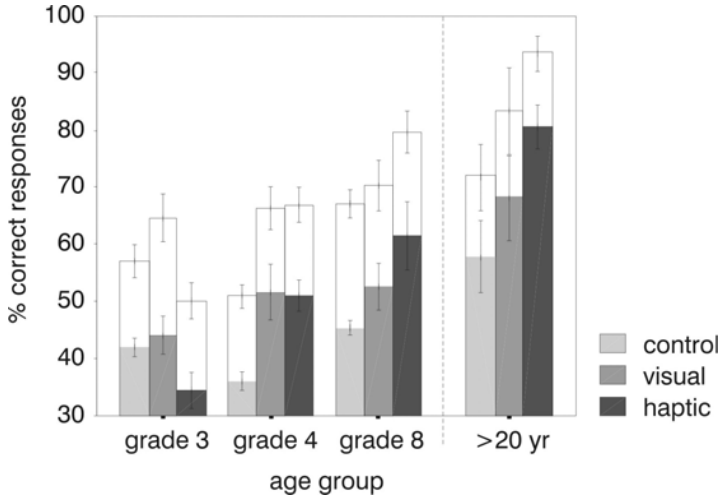


Fig. 2. Effects of prior knowledge on visual category learning for 3D objects. Average (shaded bars) and maximum (white bars) classification performance achieved during the course of learning in terms of percent correct classifications. School-children preformed up to 15 learning units, adults learned to a criterion of 90% correct. Within each age group conditions of prior knowledge were controls (light gray bars, left), visual (medium gray bars, centre), and haptic (dark gray bars, right). 5 subjects per condition; error bars S.E. (N=5).

4 Discussion

We have found a dissociation of the effects of haptic and visual prior knowledge on visual 3D object categorisation in the sense that the effect of haptic prior knowledge strongly increased in late childhood and adolescence, whereas the effect of visual prior knowledge did not. Active haptic exploration therefore provides an independent component of human 3D object recognition that, by definition, is not visual and view-dependent. We infer that humans can construct multimodal object memories from directly applying non-visual sources to visual input.

With regard of the effect of prior knowledge, we note that computer vision systems generally recognise objects and their pose in a scene by finding valid correspondences between features from an image and those of stored object models. Correspondences are said to be valid if there exists a transformation of pose, scale, and/or shape, that

maps model features onto their corresponding image features [7]. They are typically solved via different types of parallel/serial graph matching algorithms. We propose that, for human object recognition, prior knowledge facilitates the search for valid correspondences between internalised representations and input images. This may be mediated by the generation of neuronal data structures that constrain the search for valid correspondences via similar classes of procedures identified in the machine vision literature.

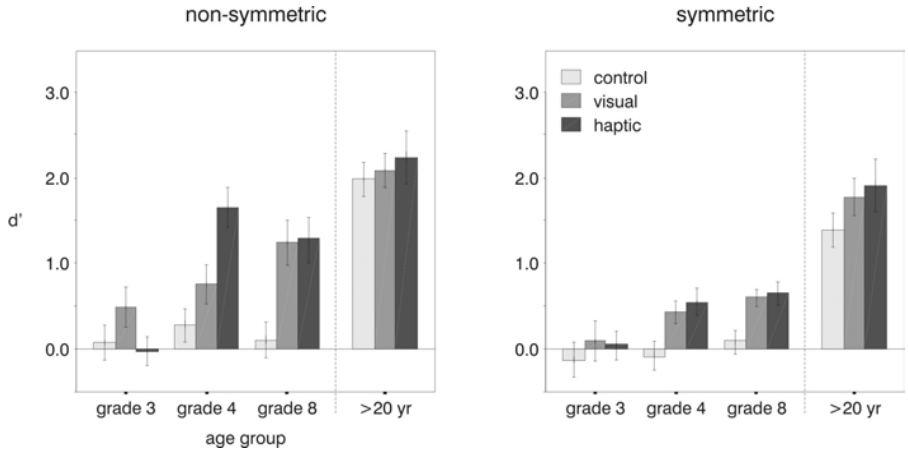


Fig. 3. Generalisation to novel views of the three 3D objects shown in Fig. 1. Bias-free measures of performance in terms of d' obtained from signal detection theory. Two decisions per subject and test view. Subjects and conditions of prior knowledge as in Fig. 2

The construction of multimodal representations of object recognition under the influence of haptic prior knowledge further raises the question of whether they are object-centred or view-dependent. We argue that prior knowledge facilitates the solution of the correspondence problem for visual object recognition independently of the structure of indexing primitives *per se*. Indeed, technical object recognition paradigms vastly differ in the complexity of indexing primitives but all require the solution of the correspondence problem. The choice of indexing primitives, ranging from 2D points to 3D volumetric primitives, depends on the nature of the data base [8]. This suggests that in human object recognition the complexity of indexing primitives is also task-dependent and not a fixed characteristic as such. However it is implausible that prior knowledge from haptic exploration should be encoded in terms of image features. Thus we conclude that the indexing primitives of resulting multimodal representations also range in complexity somewhere between 2D features and 3D primitives.

References

1. Biederman, I.: Human image understanding: recent research and a theory. Computer Vision, Graphics, and Image Processing 32, (1985) 29-73
2. Bülthoff, H.H. & Edelman, S.: Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proc. Natl. Acad. Sci. USA 89 (1992) 60-64
3. Liu, Z.: Viewpoint dependency in object representation and recognition. Spatial Vision 9 (1996) 491-521
4. Caelli, T., Johnston, M. & Robison, T.: 3D object recognition: Inspirations and lessons from biological vision. In: Jain, A.K., Flynn, P. (eds.): Three-dimensional object recognition systems. Elsevier, Amsterdam, (1993) 1-16
5. Rentschler, I. & Caelli, T. Visual representations in the brain: Inferences from psychophysical research. In: H. Haken, M. Stadler (eds.): Synergetics of Cognition. Springer-Verlag, Berlin, (1990) 233-248
6. Osman, E., Pearce, A., Jüttner, M. & Rentschler, I.: Reconstructing mental object representations: A machine vision approach to human visual recognition. Spatial Vision 13 (2000) 277-286
7. Grimson, W.E., Lozano-Pérez, T., White, S.J. & Noble, N.: Recognizing 3D objects using constrained search. In: Jain, A.K., Flynn, P. (eds.): Three-dimensional object recognition systems. Elsevier, Amsterdam (1993) 259-284
8. Dickinson, S. Part-based modeling and qualitative recognition. In: Jain, A.K., Flynn, P. (eds.): Three-dimensional object recognition systems. Elsevier, Amsterdam, (1993) 201-228

Author index