From: José Mira-Mira, editor, *Proceedings of the International Conference on Brain Processes, Theories and Models*. MIT Press, November 1995.

BUILDING VISUAL CORRESPONDENCE MAPS — FROM NEURONAL DYNAMICS TO A FACE RECOGNITION SYSTEM

Rolf P. Würtz

Computing Science Dept., University of Groningen,
P.O. Box 800, 9700 AV Groningen, The Netherlands
Ph.: +31 50 636496, fax:+31 50 633800, email: rolf@cs.rug.nl

Abstract

On the basis of a pyramidal Gabor function representation of images two systems are presented that build correspondence maps between presegmented memorized models and retinal images. The first one is formulated close to the biology of neuronal layers and dynamic links. It extends earlier ones by a hierarchical approach and background independence. The second system is formulated in a way that is efficiently implementable on digital computers but captures the crucial properties of the first one. It has the capability for object recognition under realistic circumstances which is demonstrated by recognizing human faces independently of their hairstyle.

1 Introduction

A major conceptual task for developing theories about brain function is to decide on the level of detail that is to be modeled vs. the level of abstraction necessary to capture essential properties of higher brain functions. This paper shows an attempt to do modelling on two quite different levels while keeping the properties considered crucial.

We will present two systems that can solve the *visual correspondence problem* between a memorized model and a retinal image. The task is to decide which pairs of points from image and model belong together, or to the same point on the physical object that gave rise to both. We consider such a system crucial for object recognition, especially for deformable objects.

If model and image are described by local features, these features are usually very ambiguous, i.e. a given feature can typically come from various points in the object. Correspondences found on the basis of the similarity of local features must therefore be evaluated by taking their rough relative positions into account. The easy way out of the problem, namely using more global features like, e.g., components of the amplitude spectrum is blocked by the fact that these are extremely unstable under distortions or changes in the background.

Figure 1: Sampling on the different frequency levels. The black spots mark the centers of the wavelet kernels for different spatial frequencies. The upper row shows the representation of the image, the lower one those of the model. Missing points have been eliminated due to either background influence or very low response amplitude. The neuronal systems has been simulated for the lower two levels (leftmost 2 columns) only. The neuronal layers are rectangular, those neurons without a location in the model representation are part of the layer dynamics but do not make or receive dynamic links. All three layers have been used in the template matching scheme.

In section 2 we outline the representation of visual information which is based on the Gabor function model of simple cells in the primary visual cortex. Section 3 presents a detailed neuronal model of a system that solves the visual correspondence problem by an active process that starts with the ambiguous feature similarity map and sorts out the correspondences that also show the correct arrangement. It is based on the Dynamic Link Architecture [5, 6, 4] and extends earlier systems by a shorter serial processing component and independence of the background. This system is modeled by a set of differential equations that govern the dynamics of neuronal layers and the dynamic links between them.

The successful simulations of this systems are, however, too computationally intensive to test a recognition system under realistic circumstances. Therefore, in section 4 we replace the dynamics by a hierarchical template matching scheme. This scheme can be accompanied by a phase matching component that can add subpixel accuracy to the map-

pings. How that part of the matching can be achieved in neural architecture is currently unclear.

Finally, the recognition capabilities will be demonstrated by presenting the recognition rates of human faces independent of their hairstyle.

2 Representation of models and images

The processing of a retinal grey-level image I in the primary visual cortex can be modeled by a wavelet transform based on complex-valued Gabor functions with an extra term that removes their DC-component:

$$\begin{aligned} (\mathcal{W}I)\left(\vec{k},\vec{x_0}\right) &:= \left(\psi_{\vec{k}} * I\right)(\vec{x_0}) \,, \\ \psi_{\vec{k}}(\vec{x}) &= \frac{\vec{k}^2}{\sigma^2}\exp\left(-\frac{\vec{k}^2\vec{x}^2}{2\sigma^2}\right)\left[\exp\left(i\vec{k}\vec{x}\right)-\exp\left(-\sigma^2/2\right)\right] . \end{aligned}$$

The single wavelet is parameterized by its spatial frequency $\vec{k}$, a two-dimensional vector described by length and orientation. The responses of all spatial frequencies of some fixed length form a frequency level, which assigns a small feature vector to all image points on an appropriate sampling grid (see figure 1). The components of the feature vectors correspond to the various orientations of the spatial frequency.

This pyramidal arrangement has the advantage that all responses which are influenced by the background can be discarded (see figure 1 c) and d)). The stored model (or prototype) is segmented and its representation contains only the responses of Gabor functions whose receptive fields fall completely inside the segmented area. The image representation consists of a full pyramid.

Furthermore, the required sorting out of correspondences with the wrong geometrical arrangement is greatly simplified, because it can be started on the lowest frequency level with few points, and this rough information can then be refined on higher levels, where parts of the layers can be worked on in parallel.

3 Hierarchical dynamic link matching

For the matching dynamics each frequency level of image and model is assigned a pair of neuronal layers. Each pair of layers is interconnected reciprocally by dynamic links. The layer dynamics have the general form:

$$\begin{aligned} \tau_a \frac{d}{dt}a(\vec{x}) &= -a(\vec{x}) + c_\kappa\left(\kappa(\vec{x}) - c_g\right) * \vartheta(a(\vec{x})) + c_c - c_h h(\vec{x}) + c_s s(\vec{x}) + c_\xi \xi \\ \frac{d}{dt}h(\vec{x}) &= \begin{cases} \tau_{s+}^{-1}\left(a(\vec{x}) - h(\vec{x})\right) & : \quad a(\vec{x}) > 0 \\ \tau_{s-}^{-1}\left(a(\vec{x}) - h(\vec{x})\right) & : \quad a(\vec{x}) \le 0 \end{cases} \end{aligned}$$

On the lowest frequency level the decision must be made which part of the image matches the stored model best. Both layers are wired with short-range excitation and global inhibition. In the presence of noise this supports a stable state with only one

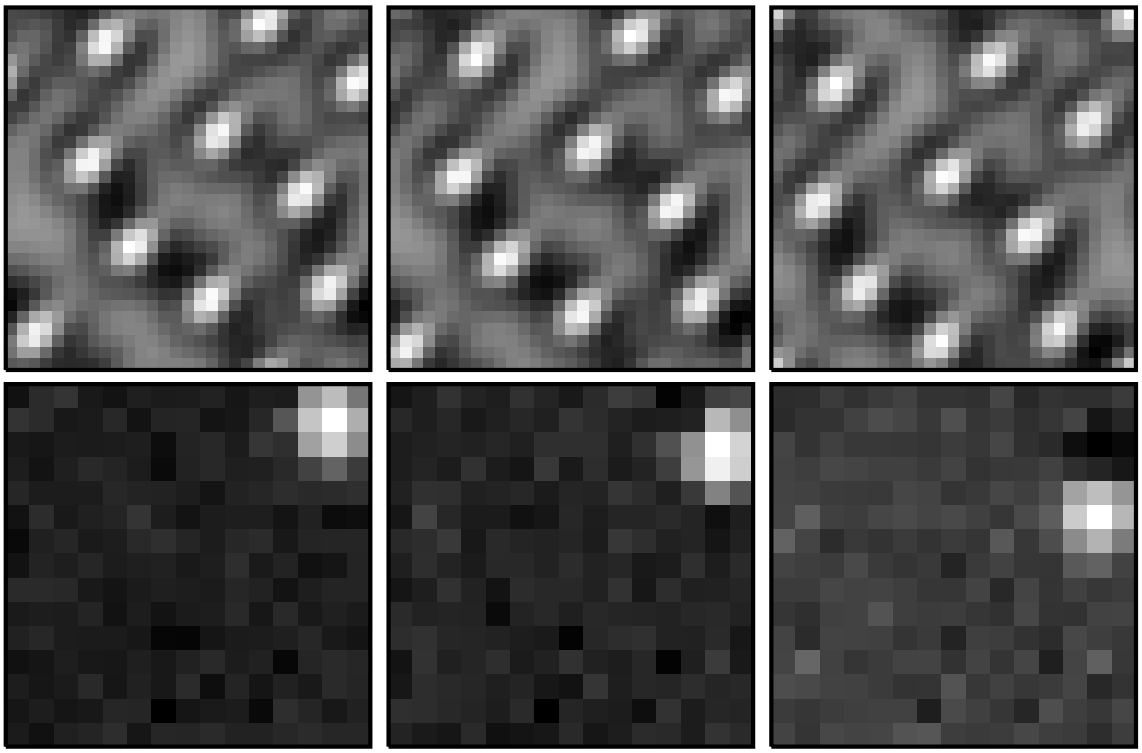

Figure 2: Layer dynamics on level 1 and 0. As a visualization of the dynamic activity on layer this figure shows three snapshots of a moving blob spaced by 25 simulation time steps (layer 0, below) and three snapshots of the multiple blobs spaced by 4 time steps (layer 1, above).

---

connected activity region (blob) [1]. A system of delayed self-inhibition ($h(\vec{x})$) is used to make this blob move across the layer.

The dynamic links between the layers are initialized to the similarities between the local features, namely vectors of Gabor amplitudes. They grow with a rate proportional to the feature similarity plus the product of the output values of the pair of neurons they connect (i.e. $\mathrm{Corr}(\vec{x},\vec{y})$ without the part in square brackets). The growth is constrained by thresholds for the total strength of outgoing and of incoming links, respectively. The link dynamics take the form:

$$\begin{aligned}
\tau_W \frac{d}{dt} W(\vec{x},\vec{y}) &= W(\vec{x},\vec{y})\mathrm{Corr}(\vec{x},\vec{y})\,, \\
\mathrm{Corr}\,(\vec{x},\vec{y}) &= \vartheta\left(a\left(\vec{x}\right)\right)\vartheta\left(a\left(\vec{y}\right)\right) + c_{\mathcal{S}}\mathcal{S}\left(\vec{f}\left(\vec{x}\right),\vec{f}\left(\vec{y}\right)\right) \\
&\quad \left[+ c_{low}\Theta\left(W\left(\vec{x}_{low},\vec{y}_{low}\right) - c_t\right)\vartheta\left(a\left(\vec{x}_{low}\right)\right)\vartheta\left(a\left(\vec{y}_{low}\right)\right)\right] \\
\int W(\vec{x},\vec{y})\,d^2x &\leq 1\,, \quad \int W(\vec{x},\vec{y})\,d^2y \leq 1
\end{aligned}$$

In the beginning both blobs move freely and independently on their corresponding

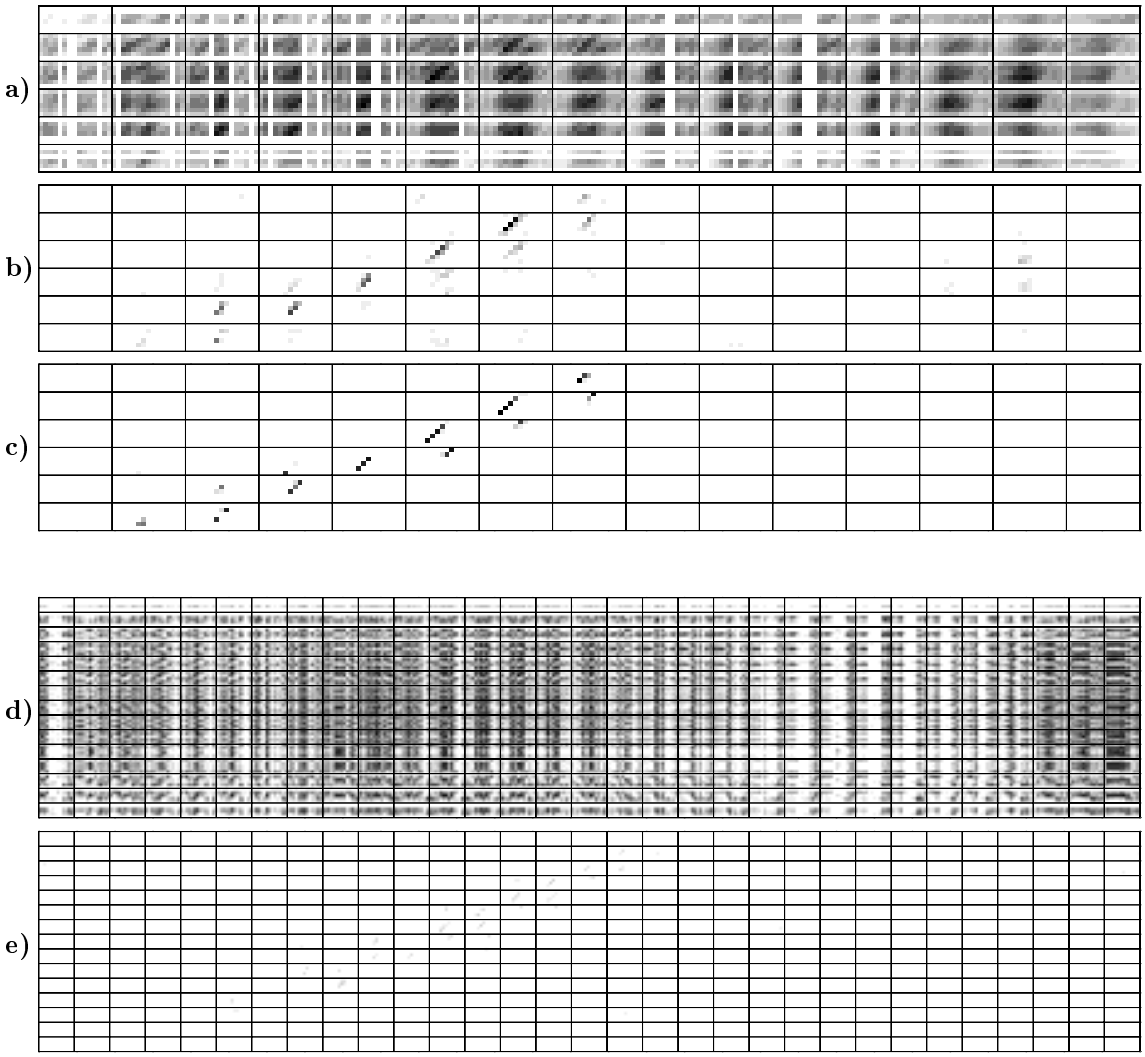

Figure 3: The development of the dynamic links. Each little rectangle contains the link strengths between one horizontal scan line in model and image, respectively. Ideal correspondences (e.g. for identical images) would show no links besides black diagonals in the rectangles belonging to corresponding lines. **a):** In the beginning the links on the lowest level reflect only the feature similarities, which are highly ambiguous. **b):**After 390 time steps the dynamics on this level have sorted out the correct correspondences, and the first links have grown above the threshold where they are allowed to influence the links on the **d)**higher frequency level. In that level link strengths are still proportional to the feature similarities, and the ambiguities are even worse. In the bottom figures **c)** and **e)** (snapshots after 1000 time steps) the links on both frequency levels are restrained to the correct correspondences, the remaining ambiguities are due to the coarse sampling, as the true correspondence would fall between sampling points. This problem can be solved by matching the Gabor phases in the template matching scheme, but has currently no neuronal implementation.

layers (see figure 2, lower part). Correlations make some links between the layers grow, others decay. After some time, the links have become strong enough that the image blob can only exist inside the region which corresponds to the model. From then on, the blob decays outside this region after a while and spontaneously reforms inside the region. When the links have grown even stronger, the image blob does not leave the region any more, and the correct links grow until a one-one mapping has been reached.

The layer dynamics on the higher levels have mexican-hat-type interaction with a kernel whose maximum is slightly off center. With appropriate parameters these dynamics support a structure of many small blobs moving coherently across the layer. The delayed self-inhibition is not active on the higher levels ($c_h = 0$) See the upper part of figure 2 for a visualization of these dynamics.

The link dynamics on a higher level are triggered once some link on the previous level has reached a threshold. Their growth rates have the same form as on the lowest level, with the extra term in square brackets, which supports only the connections that have already strong links on the previous layer (Θ denotes the Heaviside function).

With the feature similarities taken from images of human faces these dynamics can establish rough correspondences on the lowest level, which are then refined on the higher ones. This has been demonstrated by simulation of the first two levels. See figure 3 for the development of the links on the lowest and the next level. The multitude of blobs on the higher levels allows a partly parallel refinement of the correspondences estimated on lower ones. As the processing time of the single blob dynamics increases like n^4 with the (linear) layer size this system is inherently faster than others that use only one pair of layers [3, 7, 2].

4 Hierarchical template matching

For a faster implementation the following simplifications to the neuronal systems have been made.

The *coarse localization* of the counterpart of the model in the image is done by global template matching of the vectors of Gabor amplitudes on the lowest frequency level. This is not very expensive, because the resolution is low, and yields a first rough correspondence mapping. It simulates the effect of dynamic link matching with a single blob, although now only one link per model point can exist.

Consequently, the *mapping refinement* is done by local template matching with amplitudes from the next higher frequency level. For this, the model is split up into several patches that independently search for correspondences in an area defined by the coarse mapping already known.

If some feature similarities are very poor the situation may occur that no link significantly wins over the others. This will be definitely the case when occlusion occurs. In order to simulate this here, such point pairs are simply dropped from the mapping. This results in a mapping with holes, but leads to more reliable correspondences.

Mappings acquired using only the amplitudes of the Gabor responses are not very precise, because the fine geometrical information resides in the phases. On the other hand, the phases or the full complex responses are not suitable for template matching (or

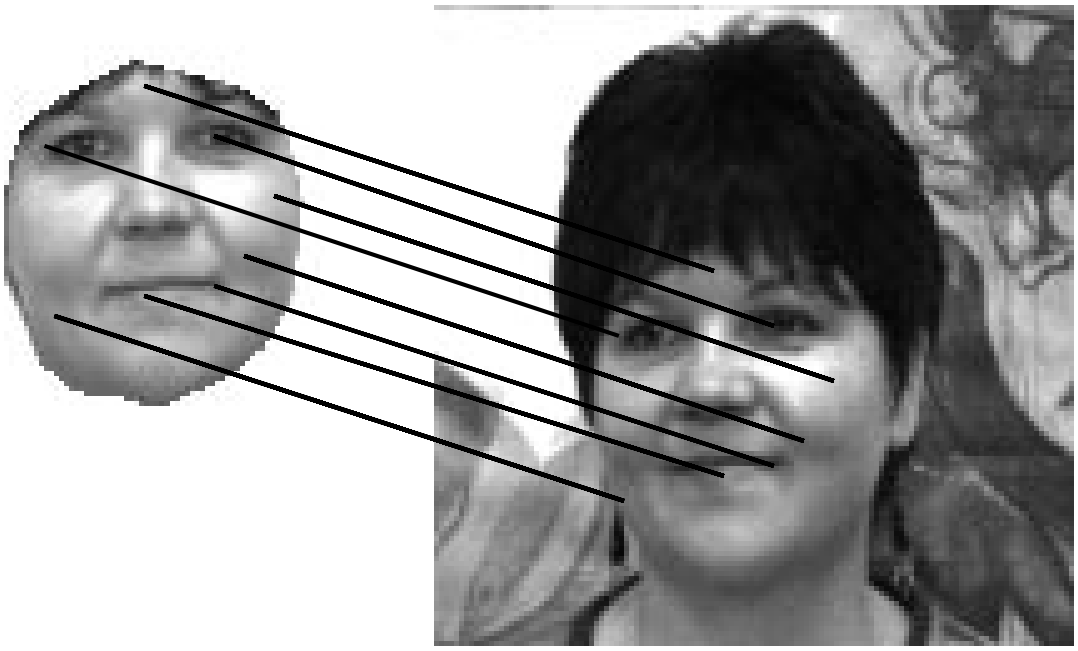

Figure 4: Correspondences on the highest level. This figure shows selected correspondences from the mapping obtained by hierarchical template matching on the highest frequency level. Although the refinement steps for distant locations are independent of each other, the hierarchy suffices to rule out false correspondences. Due to the phase matching part these correspondences are very accurate.

---

dynamic link matching) because they depend strongly on the sampling grid. Therefore, a *local phase matching* has been implemented that enhances the accuracy of amplitude-based mappings. This can be done in parallel on all model points. It has currently no counterpart in the neuronal system.

The correspondences in figure 4 are chosen from the mapping on the highest frequency level. Their accuracy shows the success of this scheme. Although the relative arrangement of the mapping points was transported up from the lowest frequency level no grossly false correspondences have been introduced. For more details on the mapping quality see [8].

5 Recognition

The procedures just described yield a correspondence map on every frequency level (see figure 1). Each of them can be used for recognition in the following way. An incoming image is matched to every model out of a database. Some global similarity is calculated from the actual feature similarities and the distortion of the mapping. The model with the highest similarity is the recognized one; the recognition is *significant* if the distance of the highest similarity to the distribution of all similarities exceeds a suitable threshold. This scheme has already been applied successfully in [4].

Method	M1 ↔ I1		M1 ↔ I2		M2 ↔ I1		M2 ↔ I2	
	C	**S**	**C**	**S**	**C**	**S**	**C**	**S**
Hier. level 0	71	54	68	23	42	17	41	11
Hier. level 1	40	29	62	45	73	54	59	18
Hier. level 2	15	12	20	8	24	23	50	23
Hier. total	99	95	93	76	99	94	85	52
Level 2 only	94	89	86	69	94	93	78	49
FACEREC	95	93	92	81	19	1	14	1

Table 1: Recognition Results. This table shows the performance of the single hierarchy steps (levels 0,1,2), the total hierarchy, the highest level alone and FACEREC, an earlier system without mechanisms for background independence. Model base **M1** contains the faces without segmentation, **M2** the same ones with the hair removed. The **C** columns show the number of correct recognitions, the **S** the significantly correct ones. All numbers are percentages of the number of test images.

In the coarse-to-fine hierarchical matching procedure described above the notion of significance can been exploited to build a hierarchical recognition scheme: If a recognition is significant on one frequency levels, the higher ones need not be evaluated.

In order to compare the background independence with the FACEREC algorithm from [4] two model databases have been set up: **M1** with model segments that were rectangular and uniform in size and **M2** with model segments that were created by hand and excluded the hair from the face images. Thus the person's hair was treated as background, there face proper as object. Both model bases contained 83 persons looking straight into the camera.

For the experiments in table 1 two different image databases have been used: **I1** containing the same persons looking 15° to their side, which can be considered as a moderately difficult case compared to **I2** which contained **I1** and for each person one image with the head turned by 30° and one with a different expression. For each image database, the significance thresholds have been adjusted such that no false positive recognitions occurred. The important performance measure is thus the number of significant recognitions.

Table 1 shows that the total recognition performance of the hierarchical scheme is scheme is better than that of the highest frequency level alone. Thus the distinguishing features are distributed over the scales in a way which is not yet completely understood. The FACEREC system described in [4], which has no mechanisms for background independence, performs about as well as the hierarchical system with model database **M1**, but breaks down completely with **M2**, where the background dependence of the feature vectors becomes serious.

6 Discussion

Two models have been proposed that are capable of solving the visual correspondence problem and are to a certain extent neurobiologically plausible. Although full experimental evidence for the existence of dynamic links in the brain is not yet available there are good theoretical reasons to postulate some binding mechanism between pairs (or groups) of neurons [5, 6, 4]. This scheme is an ideal framework for finding correct point correspondences. Its partly sequential nature poses problems with the processing times under realistic conditions. These problems can be greatly alleviated by the hierarchical approach presented here.

Then the neuronal model has been simplified in order to allow rapid simulation on a workstation. It could be shown that a face recognition system based on this model performed reasonably and was superior to an earlier system in the presence of structured background.

This paper shows a successful approach of modeling at two levels of abstraction at the same time. Both models have influenced each other by giving ideas and posing new open questions such as what might be the neuronal analogue to the phase matching in the simplified model.

Acknowledgements

The author is greatly indebted to Christoph von der Malsburg for excellent working conditions at the Institut für Neuroinformatik at Bochum, Germany, fo the whole concept of dynamic link matching and for a multitude of ideas which have shaped this work. Currently, the author's work is supported by a grant from the HCM program of the European community. Thanks also go to the German Minister for Science and technology for funding from the NAMOS project.

References

[1] S. Amari. Dynamical stability of formation of cortical maps. In M.A. Arbib and S. Amari, editors, *Dynamic Interactions in Neural Networks: Models and Data*. Springer, 1989.

[2] W. Konen and J.C. Vorbrüggen. Applying dynamic link matching to object recognition in real world images. In S. Gielen, editor, *Proceedings of the International Conference on Artificial Neural Networks*. North-Holland, Amsterdam, 1993.

[3] Wolfgang Konen, Thomas Maurer, and Christoph von der Malsburg. A fast dynamic link matching algorithm for invariant pattern recognition. *Neural Networks*, 7(6/7):1019–1030, 1994.

[4] Martin Lades, Jan C. Vorbrüggen, Joachim Buhmann, Jörg Lange, Christoph von der Malsburg, Rolf P. Würtz, and Wolfgang Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.

[5] Christoph von der Malsburg. The correlation theory of brain function. Technical report, Max-Planck-Institute for Biophysical Chemistry, Postfach 2841, Göttingen, FRG, 1981. Reprinted 1994 in: Schulten, K., van Hemmen, H.J. (eds.), *Models of Neural Networks*, Vol. 2, Springer.

[6] Christoph von der Malsburg. Nervous structures with dynamical links. *Ber. Bunsenges. Phys. Chem.*, 89:703–710, 1985.

[7] Laurenz Wiskott and Christoph von der Malsburg. Dynamic link matching with running blobs. In preparation, 1994.

[8] Rolf P. Würtz. *Multilayer Dynamic Link Networks for Establishing Image Point Correspondences and Visual Object Recognition*, volume 41 of *Reihe Physik*. Verlag Harri Deutsch, Thun, Frankfurt am Main, 1995.