Neural Networks 41 (2013) 137–146

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2013 Special Issue

Learning invariant face recognition from examples

Marco K. Müller, Michael Tremer, Christian Bodenstein, Rolf P. Würtz *

Institut für Neuroinformatik, Ruhr-Universität, D-44780 Bochum, Germany

ARTICLE INFO

Keywords:
Rank statistics
Learning invariance
Face recognition
Spike time
Spiking neural network
Controlled generalization
Situation independence

ABSTRACT

Autonomous learning is demonstrated by living beings that learn visual invariances during their visual experience. Standard neural network models do not show this sort of learning. On the example of face recognition in different situations we propose a learning process that separates learning of the invariance proper from learning new instances of individuals. The invariance is learned by a set of examples called model, which contains instances of all situations. New instances are compared with these on the basis of rank lists, which allow generalization across situations. The result is also implemented as a spike-time-based neural network, which is shown to be robust against disturbances. The learning capability is demonstrated by recognition experiments on a set of standard face databases.

1. Introduction

Recent years have brought significant improvements in computer vision in real-world settings. Nevertheless, the performance is still plagued by the fact that the same object produces very different images in different *situations*. We use the term here in a very wide and abstract sense, such that it covers different poses, illuminations, types of camera, distance to the camera, position within the camera view, background, hairstyle, accessories worn, facial expression, and aging.

Invariant face recognition then refers to estimating the identity of a person irrespective of the situation. Human perception is excellent at both finding the identity of a known person and estimating the situation of both known and unknown persons on the basis of a facial image. ("This is John in his twenties in the disco" or "Jenny is sunbathing on the beach and seems to enjoy it".) Certainly, the human visual system is good at the *separation* of personal identity and situation. This is possible by using the vast visual experience acquired with many persons in many situations.

From a machine learning point of view, the requirement to recognize identity independent of situation is a case of generalization and should be learned *autonomously* like humans do in their early childhood.

However, invariance under even a simple visual transformation such as translation in the image plane is not a generalization performed naturally by known learning mechanisms in neural networks. Therefore, methods to *control the generalization* on the basis of examples are required.

This may seem like a contradiction. On the one hand, learning is supposed to be autonomous, but on the other hand it should be controlled. But clearly, autonomy without any control is not desirable in technical systems. Therefore, ways must be found to exert this control with minimal effort.

From the viewpoint of Autonomous Learning the challenge is to split up the learning of recognition of faces into two subsystems. One that learns the invariance on the basis of a set of examples, and another one that can learn new identities from just a single example.

Invariances can, to a limited degree, be learned from real-world data based on the assumption that temporally continuous sequences leave the object identity unchanged (Bartlett & Sejnowski, 1998; Földiák, 1991; Hinton, 1987; Wiskott & Sejnowski, 2002). Slow feature analysis has recently been successfully applied to 3D rotation by Franzius, Wilbert, and Wiskott (2011).

Nevertheless, all successful recognition systems have the required invariances built in by hand. This includes elastic graph matching (Lades et al., 1993) and elastic bunch graph matching (EBGM) (Wiskott, Fellous, Krüger, & von der Malsburg, 1997), where the graph dynamics explicitly have to probe all possible variations in order to compare an input image with the stored gallery. Neural architectures that perform this matching include (Jitsev & von der Malsburg, 2009; Lücke, Keck, & von der Malsburg, 2008; Wiskott & von der Malsburg, 1996; Wolfrum, Wolff, Lücke, & von der Malsburg, 2008), with the more recent ones being massively parallel and able to account for invariant recognition with processing times comparable to that of the visual system. These methods work fine for the recognition of identity under changes in translation, scale, and small deformations, including small changes in three-dimensional pose.

Invariances for which explicit modeling is difficult, like large pose differences or illumination changes, can be handled by elastic bunch graph matching if bunch graphs are supplied for a coarsely

* Corresponding author.
E-mail addresses: marco.k.mueller@rub.de (M.K. Müller), michael.tremer@ini.rub.de (M. Tremer), christian.bodenstein@ini.rub.de (C. Bodenstein), rolf.wuertz@ini.rub.de (R.P. Würtz).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.07.006

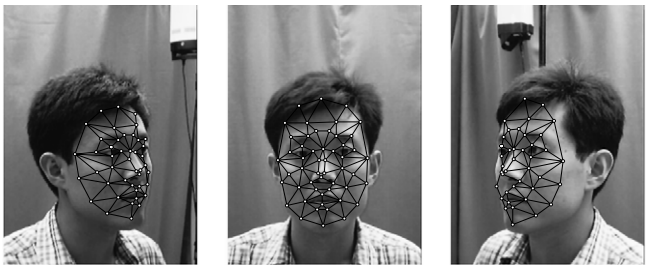

Fig. 1. Bunch graphs for different poses in the CAS–PEAL database. Images in different poses are not directly comparable because of different node numbers and strongly distorted features.

sampled set of situations, e.g., 10 different head poses (see Fig. 1). This is problematic from a technical point of view (Murphy-Chutorian & Trivedi, 2009), because for a recognition system for many persons it is infeasible to store and match all persons in all possible poses or illuminations. It is also improbable that the brain would employ such a strategy because of the same waste of memory resources.

We here present a system that can learn invariances in a moderately supervised way from a set of examples of individual faces in several situations. Person identification generalizes to other individuals that are known only in *one* situation.

The similarity we will introduce in this paper is a special case of *rank correlation*, one example being Spearman's rank order correlation coefficient (Press, Flannery, Teukolsky, & Vetterling, 1988). This sort of statistics has been used for the evaluation of biometric systems (Rukhin & Osmoukhina, 2005) and for face matching by Ayinde and Yang (2002). Here, we apply it to guide generalization into a desired direction by the presentation of examples.

2. Invariance by rank list comparison

2.1. Elastic bunch graph matching

Recognition by graph matching (Lades et al., 1993; Wiskott et al., 1997) compares a given *probe* image with *gallery graphs* G_g of all known persons. The gallery graphs consist of N nodes, which are labeled with local feature vectors $G_{g,n}$, for the probe image. The same feature types are known at all image locations and denoted with $P_{\text{full}}(\vec{x})$. Correspondences between image points are estimated in a process called landmark finding by finding the positions $\vec{x}_n^{\text{opt}}$ which maximize the similarity

$$\frac{1}{N}\sum_{n=1}^{N}\max_g S_J(P(\vec{x}_n), G_{g,n})$$

with $S_J(J_1, J_2)$ being a similarity function between two local feature vectors.

Once nodes are positioned correctly, the graph P representing the probe image contains N nodes with jets $P(\vec{x}_n^{\text{opt}})$. For recognition, a similarity between persons is calculated by averaging local similarities $S_J(P, G_g, n)$ of *corresponding* features. The local similarity function need not be identical to the one used for landmark finding. The topology of all graphs is the same and only relevant for landmark finding. Finally, the recognized identity is G_g with

$$g = \arg\max_g \frac{1}{N}\sum_{n=1}^{N} S_{\text{loc}}(P, G_g, n). \tag{1}$$

Collections of graphs with the same topology and with mutually corresponding features are referred to as bunch graphs, because they can be interpreted as a single graph containing the bunches of features from different images. The whole gallery can thus be seen as one bunch graph $\mathcal{G}$.

2.2. Local features and similarity functions

This general algorithm can be run with different types of local feature vectors and similarity functions. We have shown elsewhere (Günther & Würtz, 2009) that this choice significantly influences the recognition rate. The method introduced here can be used with arbitrary feature types and similarities. In this paper we use two different types of local features, namely *Gabor jets* (following Lades et al., 1993; Wiskott et al., 1997) and *Local Gabor Binary Pattern Histogram Sequences (LGBPHS)*, which show very good performance in face recognition (Zhang, Shan, Gao, Chen, & Zhang, 2005).

2.2.1. Gabor jets

Mean-free Gabor wavelets are widely used in face recognition (Lades et al., 1993; Wiskott et al., 1997). With the center frequency $\vec{k}$ as the parameter, they take the form:

$$\psi_{\vec{k}}(\vec{x}) = \frac{\vec{k}^2}{\sigma^2} e^{-\frac{\vec{k}^2\vec{x}^2}{2\sigma^2}} \left(e^{i\vec{k}\vec{x}} - e^{-\frac{\sigma^2}{2}}\right). \tag{2}$$

A typical parameterization for face recognition employs a family of $K = 40$ Gabor wavelets $\psi_{\vec{k}_j}$ $(j = 1, \ldots, K)$ at 5 scale levels and 8 directions. The convolution of an image with a Gabor wavelet $\psi_{\vec{k}_j}$ at image position $\vec{x}$ results in a complex-valued response, which can be split into *amplitudes* a_j and *phases* ϕ_j as $a_j \cdot e^{i\phi_j}$. The responses from all Gabor wavelets taken at the same position $\vec{x}$ in the image are called a *Gabor jet* $\mathcal{J}$, which codes the texture information around the offset point.

For Gabor jets with amplitudes a_j, the following similarities are employed following Gonzáles et al. (2007), Günther and Würtz (2009) and Wiskott et al. (1997), respectively:

$$S_{\text{Abs}}(\vec{J}_1, \vec{J}_2) = \frac{\sum_{j=1}^{K} a_{1,j} \cdot a_{2,j}}{\sqrt{\left(\sum_{j=1}^{K} a_{1,j}^2\right)\left(\sum_{j=1}^{K} a_{2,j}^2\right)}}. \tag{3}$$

$$S_{\text{Canb}}(\vec{J}_1, \vec{J}_2) = K - \sum_{j=1}^{K} \frac{|a_{1,j} - a_{2,j}|}{\max(|a_{1,j}| + |a_{2,j}|, 10^{-6})}. \tag{4}$$

$$S_{\text{Manh}}(\vec{J}_1, \vec{J}_2) = -\frac{\sum_{j=1}^{K} |a_{1,j} - a_{2,j}|}{\sum_{j=1}^{K} |a_{1,j}| \sum_{j=1}^{K} |a_{2,j}|}. \tag{5}$$

2.2.2. LGBPHS

The features of *Local Gabor Binary Pattern Histogram Sequences (LGBPHS)* are built from Gabor amplitudes with the same parameters as above. At point $\vec{x}$ and for center frequency $\vec{k}_j$, the Local Binary Pattern (LBP) is a binary number calculated from the amplitudes at the 8 neighboring pixels $\vec{x}_p$:

$$\text{LBP}_j = \sum_{p=0}^{7} \text{Bool}\left(a_j(\vec{x}_p) \geq a_j(\vec{x})\right) \cdot 2^p \tag{6}$$

where $\text{Bool}(\cdot)$ yields the binary truth value of its argument.

These LBPs are histogrammed into 16 bins over a 10×10 pixel region around $\vec{x}$ for each center frequency $\vec{k}_j$ yielding histograms H_j. These are local features of the point $\vec{x}$, which are compared by the sum of the minima of their components. Finally, the similarity is added over all center frequencies:

$$S_{\text{LGBPHS}}(H^1, H^2) = \sum_{j=1}^{K}\sum_{i=1}^{16} \min\left(H_j^1(i), H_j^2(i)\right). \tag{7}$$

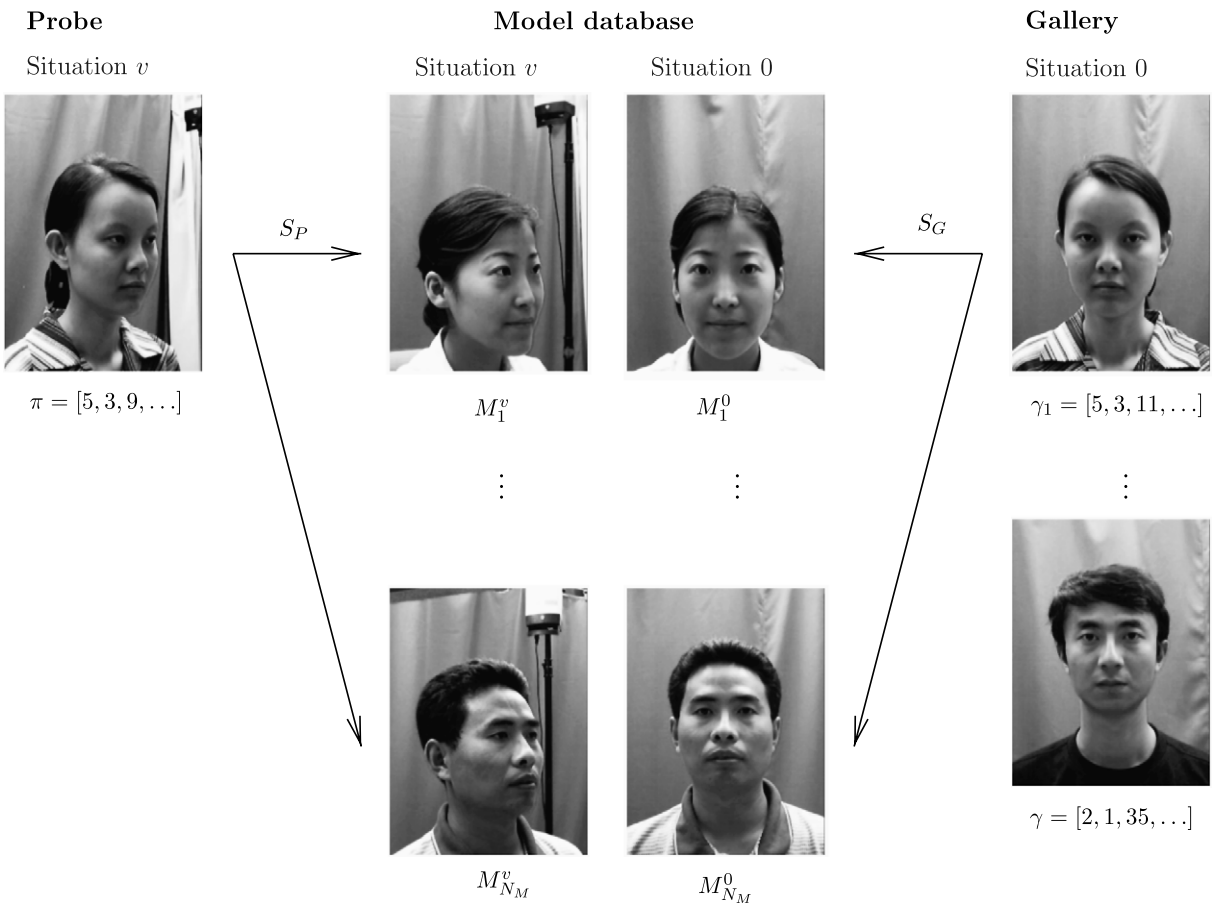

Fig. 2. Situation-independent recognition is mediated by a model database of *some* persons in *all* situations (of which only two are shown). Each node of the probe and gallery graphs is coded into rank lists π and γ by their similarities to the models. These rank lists are comparable, while the graph similarities are not. Note that the probe person is not in the model database. (Node indices have been dropped for clarity, and the numbers in the rank lists are just examples).

2.3. How to compare images in different situations

The graph matching procedure assumes that the same or very similar local features can be found in the probe image as well as in the gallery images. This is only true for translation in the image plane and small distortions. For other changes like in-plane rotation or rescaling, the expected variations can be modeled and become part of the similarity function (Günther & Würtz, 2009).

For the recognition of an arbitrary subject a large *gallery* database is created, which contains all known subjects in a preferred situation $v = 0$. Throughout this paper, this situation will be a frontal pose under frontal illumination. This choice has also been shown to be favorable by Müller, Heinrichs, Tewes, Schäfer, and Würtz (2007).

The technique rests on the heuristic that persons that are similar in one situation will also be similar in another one. The similarities themselves may vary widely, but the order of similarities to different persons should be maintained. The variations between situations are modeled by a database of *models*, which are known in all situations, and personal identity is coded by a similarity rank list to the models of the same situation (see Fig. 2 for an illustration). The collection of graphs M_m^v for all models of a single situation constitutes a bunch graph $\mathcal{M}^v$.

The rank list for a probe subject P is created as follows. First, all local similarities S^v to all model images M_m^v are calculated. For each index n and situation v a rank list ρ_n^v is created, which contains the rank of similarity for each model index m, so that for each pair of model images M_m^v, $M_{m'}^v$ the following holds ($\rho_n^v(m) \in \{1, \ldots, N_M\}$):

$$\rho_n^v(m) < \rho_n^v(m') \Rightarrow S_{\text{loc}}(P, M_m^v, n) \geq S_{\text{loc}}(P, M_{m'}^v, n). \tag{8}$$

The most similar model candidate would be the one with $\rho_n^v(m) = 0$, the follower-up the one with $\rho_n^v(m) = 1$, etc. For each node n, these lists now serve as a representation of the probe graph P. For varying P we will use the notation $\rho_n^v(P, m)$.

Each subject G_g in the gallery is assigned a rank list representation $\gamma_{g,n}$ by matching each of its landmarks to those of the model subjects in the preferred situation:

$$\gamma_{g,n}(m) = \rho_n^0(G_g, m), \quad m = 1 \ldots N_M. \tag{9}$$

For recognition we first assume that a *probe* image P^v appears in the *known* situation v. The requirement to know the situation will be dropped in Section 2.7. This probe is also represented as a similarity rank list π_n^v for each landmark of all models in situation v:

$$\pi_n^v(m) = \rho_n^v(P^v, m), \quad m = 1 \ldots N_M. \tag{10}$$

2.4. Rank list comparison

Having represented gallery and probe graphs by a set of rank lists of equal length all that is required for invariant recognition is a function $S_{\text{rank}}(\pi, \gamma)$ that measures the similarity of these rank lists. Such a function should fulfill three requirements:

1. It should take values between 0 and 1 and be maximal for two identical rank lists:

$$\forall \rho_1, \rho_2 : 1 = S_{\text{rank}}(\rho_1, \rho_1) \geq S_{\text{rank}}(\rho_1, \rho_2); \tag{11}$$

2. It should be high if many model indices appear at the same rank and low if the ranks are mixed;

3. Cooccurrences with high image similarities (i.e., with low rank ρ) should be weighted more strongly than those with low ones. It is expected that high image similarities are more informative about identity.

These requirements are fulfilled if $S_{\text{rank}}(\rho_1, \rho_2)$ takes the following form:

$$S_{\text{rank}}(\rho_1, \rho_2) = \frac{1}{F}\sum_{m=1}^{N_M} f(\rho_1(m) + \rho_2(m)) \tag{12}$$

$$F = \sum_{m=1}^{N_M} f(2m) \tag{13}$$

where f is a monotonically decreasing function and F a normalization factor, which enforces the maximal similarity of 1.

From the comparison with Spearman rank-correlation one would expect the difference between the individual ranks rather than the sum. The minus sign, however, would weigh high entries, which correspond to images of low similarity, equally as low entries, thus violating the third condition.

In earlier studies (Müller, 2010; Müller et al., 2007) we have used $f(x) = (x+1)^d$ with $d \in [-2, 0)$. Here, we use $f(x) = \lambda^x$, with $\lambda \in [0.9, 1)$. This faster decaying function yields slightly better recognition results (Müller & Würtz, 2009) and, additionally, allows for a natural interpretation and implementation as a neural network (see Section 3). These considerations lead to the following form:

$$S_{\text{rank}}(\pi, \gamma_g) = \frac{1}{F}\sum_{m=1}^{N_M} \lambda^{\pi(m)+\gamma_g(m)}. \tag{14}$$

2.5. Recognition

In graphs of different situations, the nodes are numbered such that corresponding landmarks have the same value of n, across all situations. Consequently, graphs in different situations have only subsets of these landmarks as nodes, the set of node indices for each situation is denoted by $\mathcal{L}^v$.

The rank list similarity can be evaluated separately for each feature, and the resulting similarities are averaged over all features shared by the graphs in both situations. Features unavailable due to self-occlusion are ignored in the accumulated similarity as well as the normalization. It is assumed that the same subset of features is available in all images in a single situation:

$$S_{\text{rec}}(g) = \frac{1}{|\mathcal{L}^v \cap \mathcal{L}^0|}\sum_{n\in\mathcal{L}^v\cap\mathcal{L}^0} S_{\text{rank}}(\pi_n^v, \gamma_{g,n}). \tag{15}$$

As usual, the recognized person is the one with the index g that maximizes this similarity (see Eq. (1)).

2.6. Combined local similarity functions

The above recognition procedure can be applied with all local similarity functions defined in Section 2.2. Besides that, the rank lists produced by different local similarity functions can be combined by averaging over the resulting rank list similarities. Thus, global similarity functions can be created for all nonempty subsets $\mathcal{S} \subseteq \{\text{Abs, Canb, Manh, LGBPHS}\}$, the rank lists are denoted by $\pi_n^{v,\mathcal{S}}$ and $\gamma_{g,n}^{\mathcal{S}}$:

$$S_{\text{rec}}^{\mathcal{S}}(g) = \frac{1}{|\mathcal{S}|}\frac{1}{|\mathcal{L}^v \cap \mathcal{L}^0|}\sum_{s\in\mathcal{S}}\sum_{n\in\mathcal{L}^v\cap\mathcal{L}^0} S_{\text{rank}}^{s}(\pi_n^{v,s}, \gamma_{g,n}^{s}). \tag{16}$$

Clearly, Eq. (15) is a special case of this if $\mathcal{S}$ contains only one element.

2.7. Estimation of the situation proper

In a realistic setting, the situation of the probe image P is, of course, unknown. There are many ways to estimate the situation, see, e.g., Murphy-Chutorian and Trivedi (2009) or Ma, Zhang, Shan, Chen, and Gao (2006) for pose. Here, we have settled for a very simple, albeit rather inefficient (Table 2) one. This is situation estimation by standard matching of bunch graphs $\mathcal{M}^v$ for all situations, and assigning the situation with the highest similarity:

$$v_{\text{est}} = \arg\max_v \frac{1}{|\mathcal{L}^v|}\frac{1}{N_M}\sum_{n\in\mathcal{L}^v}\sum_{m=1}^{N_M} S_{\text{Abs}}(P, M_m^v, n). \tag{17}$$

In the case of N_V situations, bunch graph matching leads to N_V graphs for a given probe image P. For each situation, the average similarity of that graph to all corresponding graphs of the model is calculated. The highest similarity indicates the estimated situation v_{est}, which is used instead of the known situation in the above procedure.

2.8. Learning from an unlabeled model set

Some data sets (like, e.g., FRGC (Phillips, Flynn, Scruggs, Bowyer, & Worek, 2006)) have a training set labeled with identity, but not with situation. With a slight modification of rank list construction, such training sets can also be used as a model for the transformation to be learned. The model set consists of images labeled with an identity index i and a model index d enumerating the model images belonging to each identity. When creating the rank list ρ for a given image, local similarities $S_{\text{loc}}(P, M_{d,i})$ to all model images are calculated. The rank list is built from the ranks of all identities, based on the maximal local similarity for each identity. As in the labeled case, the length of the resulting rank lists is the number of identities in the model set:

$$\rho_n(i) < \rho_n(i') \Rightarrow \max_d S_{\text{loc}}(P, M_{d,i}, n) \geq \max_d S_{\text{loc}}(P, M_{d,i'}, n). \tag{18}$$

The maximal local similarity for each identity performs automatic situation estimation for each identity. On these rank lists, recognition proceeds as above.

3. Rank list comparison by a spiking neural network

Thorpe, Delorme, and Van Rullen (2001) have proposed a neural network that can evaluate rank codes. A set of feature detectors respond to an input pattern such that the most similar detector fires a spike first. The order in which the sent spikes arrive can then be decoded by a circuit depicted in the left half of Fig. 3.

We assume a neuronal module that calculates the similarity of stored model images to the actual probe image. Each gallery subject has one representing neuron. The similarity influences the time a neuron corresponding to this subject sends a spike. The higher the similarity the earlier the spike.

The activation in response to a spike train a_j is calculated as

$$A = \sum_{j=1}^{K} \lambda^{\text{order}(a_j)} w_j \tag{19}$$

with λ determining the activity decrease per spike. The activity A is maximal if $\text{order}(a_j)$ is the same as the order of the weights, because then the largest weight gets the smallest exponent and the largest multiplier. Put the other way around, if b_j is the sequence to elicit the largest activation the weights may be chosen as

$$w_j = \frac{1}{K}\lambda^{\text{order}(b_j)}. \tag{20}$$

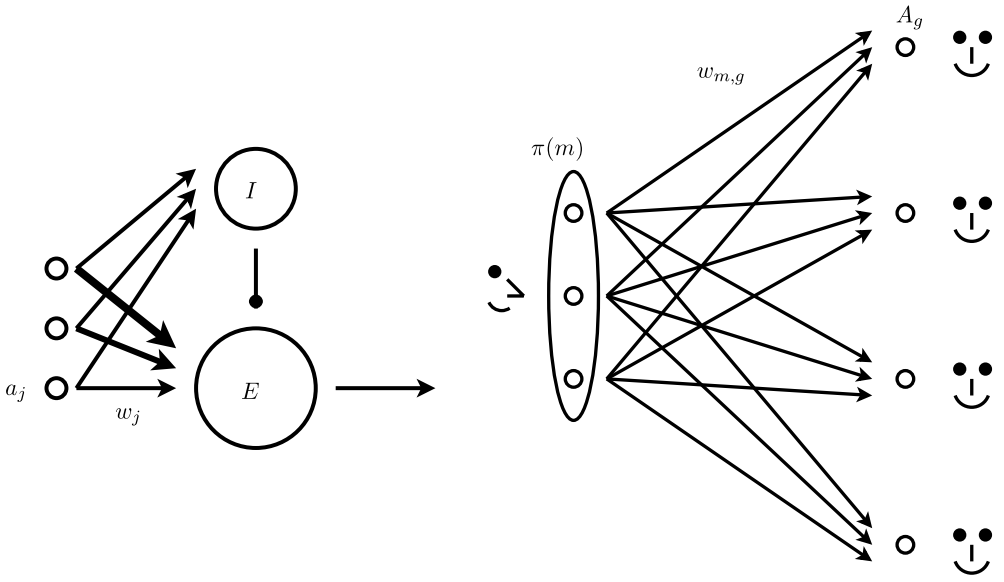

Fig. 3. Left: a neural circuit sensitive to the order of firing neurons, the preferred order is stored in the weights w_j (after Thorpe et al., 2001). Right: the same circuit is repeated for each gallery image. The probe image is represented as a rank list π according to similarities with model images in the same situation. The similarities of the gallery to the model images in neutral situation are coded in the weights $w_{m,g}$.

For the functionality only the order of the weights matters. The form in Eq. (20) is chosen for convenience and in accordance with Thorpe et al. (2001). The parameter λ needs to be adjusted, its optimal value varies with the size of the rank list N_M.

For our purposes, such a decoding circuit is required for each gallery image G_g. π is the rank list or the firing order of a number of N_M model neurons firing according to their similarity of each model image with index m to the probe image. The rank list γ_g of gallery image G_g is coded in the synaptic weights $w_{m,g}$ as follows:

$$w_{m,g} = \frac{1}{N_M}\lambda^{\gamma_g(m)}. \tag{21}$$

The activity A_g then becomes

$$A_g = \sum_{m=1}^{N_M} \lambda^{\pi(m)} w_{m,g} \tag{22}$$

$$= \frac{1}{N_M}\sum_{m=1}^{N_M} \lambda^{\pi(m)+\gamma_g(m)}. \tag{23}$$

This is precisely the similarity function between the rank lists π and γ_g introduced in Eq. (14).

We have implemented the network in a continuous-time fashion (Bodenstein, 2011), meaning that the precise spiking times are implemented as floats. This allows us to study the robustness of the network under the influence of disturbances like imprecision in spike timing, varying synaptic delays, multiple spikes, etc.

The formalization of the spiking network is as follows. After a global reset, each feature detector fires a spike at time:

$$t_i = 1 - S_{\text{loc}}(J_i^M, J_i^G). \tag{24}$$

This network has the advantage that the evaluation of situation-invariant recognition can be very fast. The similarities have values in [0, 1] and so do these spiking times. To map these values to biologically realistic timings requires a time unit of about 20 ms. This speed can only be exploited if the results are robust against noise and other disturbances of the spike time. We will analyze this property experimentally in Section 4.3.

4. Experiments and results

4.1. Face databases

4.1.1. The CAS–PEAL database

The CAS–PEAL face database (Gao et al., 2008, 2004) contains 1015 Chinese individuals in different poses and 191 in different illumination conditions. In this database, the situations are carefully controlled, which makes it the prime candidate for testing our system. We only use the poses of $-45°$, $0°$, $+45°$ and the illumination conditions as shown in Fig. 4.

4.1.2. The PIE database

The PIE database (Sim, Baker, & Bsat, 2003) contains pictures that were taken under similar pose variation for 68 identities. To test the transferability of the learned transformation from the CAS–PEAL images, the model set still contains the 500 CAS–PEAL images. Also the bunch graph is built as in the CAS–PEAL setup. Only the test images (gallery in frontal pose and probe images in $\pm 45°$) have been replaced by the PIE images. As similarities we also used S_{Abs}, S_{Canb} and S_{LGBPHS}.

4.1.3. The SCface database—pose variation

As a third database, the pose variations of the SCface database (see Fig. 5) (Grgic, Delac, & Grgic, 2009) have been used in the same way as the PIE database. This database contains 130 identities, and we picked the ones close to the poses of $-45°$, $0°$, $+45°$.

4.1.4. The SCface database—surveillance camera images

Even more interesting for real-world face recognition is the collection of poor quality surveillance camera and infrared images in the SCface database (Grgic et al., 2009) (see Fig. 5). Each setting can be interpreted as a different situation. Splitting the images into a model set (80 identities) and test set (49 identities, one identity has been used for manual labeling) enables rank list comparison for such different images as normal and infrared pictures.

4.1.5. The FRGC database

The FRGC database (Phillips et al., 2006) comes with a number of defined experiments, some dealing with 3D data. For this paper

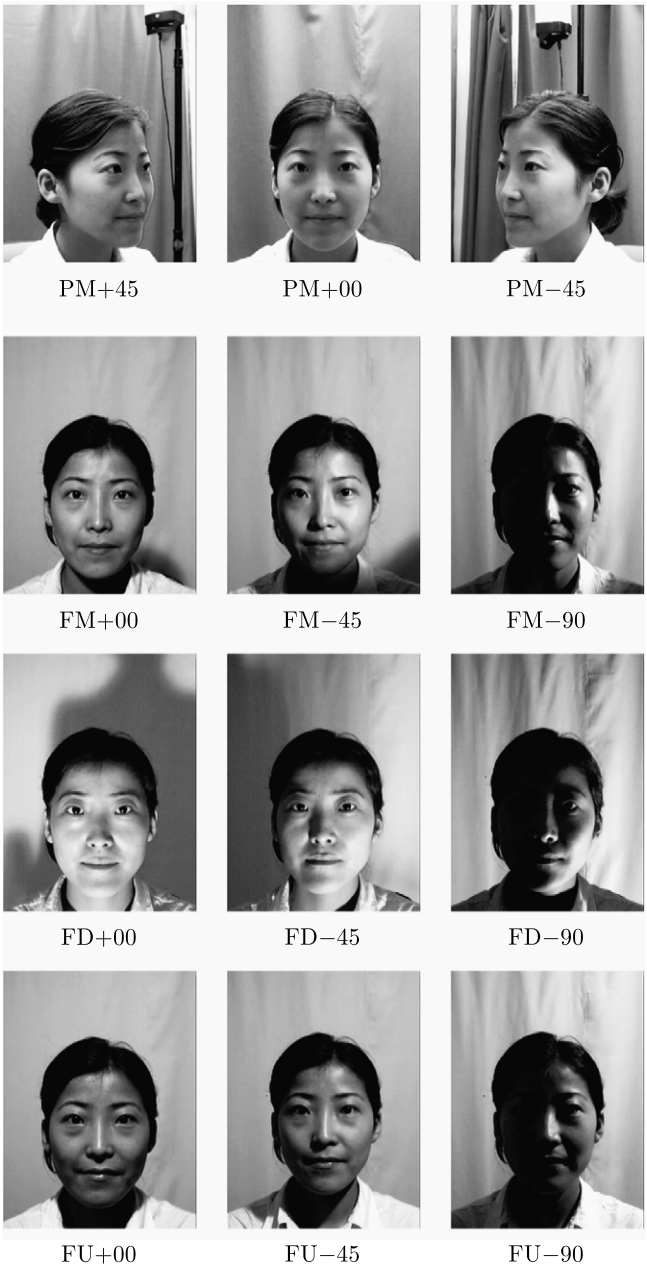

Fig. 4. Examples for pose variation (top row) and illumination variation in frontal pose from the CAS–PEAL database.

Fig. 5. Example images from the SCface database. One person is depicted by 5 surveillance cameras and two infrared cameras at different resolutions. The fourth picture in the fourth row is from a high-resolution infrared camera, the fifth a high quality portrait, which serves as gallery entry. The bottom row shows the three poses, with the middle one the gallery image.

we carried out the two experiments that use single images. Experiment 2.1 calculates similarities between all images marked as target and calculates the receiver operating characteristic (ROC) curve using a mask. Experiment 2.4 calculates similarities between target and query images. As model images we use the training images of this database. As these are only labeled with identity and not with a given situation, the rank lists are built as described in Section 2.8. As local similarities we used the combination of S_{Manh} and S_{LGBPHS}, which performed best in the case of the unmarked data.

4.2. Recognition performance of rank list comparison

The method was first tested on the CAS–PEAL face database (Gao et al., 2004). The landmarks are found by elastic bunch graph matching, starting from images of 24 subjects that were labeled by hand. From these, basic bunch graphs $\mathcal{M}^{v}_{\text{basic}}$ have been built for each situation (12 identities for pose, 8 for illumination).

The remaining 1015 subjects have been partitioned into model sets and testing sets. The testing sets provide the gallery images in the standard situation and test images for all other situations. In the pose case, we have used the first 500 subjects for model and the following 515 for testing. In the illumination case the first 100 subjects were used for model and the following 91 for testing (these are called the *standard partition* for illumination and pose, respectively).

For statistical evaluation, we have also used 100 randomized partitions with the same number of models/testing for both situations.

From the basic bunch graphs $\mathcal{M}^{v}_{\text{basic}}$ in each situation the landmarks on the model set have been determined by *incremental bunch graph building* (Heinrichs, Müller, Tewes, & Würtz, 2006; Müller et al., 2007). After EBGM was performed with $\mathcal{M}^{v}_{\text{basic}}$ on

Table 1
Recognition rates on the CAS–PEAL database for the different local similarity functions and their combinations on the standard partition. The bottom rows show recognition rates from other groups on the same database for comparison (all in %).

	Pose	Illum.
Local similarity		
S_{Abs}	97.4	82.7
S_{Canb}	95.0	76.5
S_{Manh}	97.5	81.7
S_{LGBPHS}	81.7	79.9
$S_{\text{Abs}} \diamond S_{\text{LGBPHS}}$	98.2	88.2
$S_{\text{Abs}} \diamond S_{\text{Canb}}$	98.4	83.8
$S_{\text{Abs}} \diamond S_{\text{Manh}}$	97.7	83.1
$S_{\text{LGBPHS}} \diamond S_{\text{Canb}}$	96.0	86.8
$S_{\text{LGBPHS}} \diamond S_{\text{Manh}}$	98.2	88.0
$S_{\text{Canb}} \diamond S_{\text{Manh}}$	97.9	80.9
$S_{\text{Abs}} \diamond S_{\text{LGBPHS}} \diamond S_{\text{Canb}}$	**98.9**	**88.9**
$S_{\text{Abs}} \diamond S_{\text{LGBPHS}} \diamond S_{\text{Manh}}$	98.3	87.8
$S_{\text{Abs}} \diamond S_{\text{Canb}} \diamond S_{\text{Manh}}$	98.2	84.2
$S_{\text{LGBPHS}} \diamond S_{\text{Canb}} \diamond S_{\text{Manh}}$	98.6	87.6
$S_{\text{Abs}} \diamond S_{\text{LGBPHS}} \diamond S_{\text{Canb}} \diamond S_{\text{Manh}}$	98.7	88.0
Other approaches		
Gao et al. (2008)	71.0	51.0
Zhang et al. (2008)		70.1
Tan and Triggs (2010)		72.7

Table 2
Computation times (in seconds) on a 3.4 GHz Xeon for the components of the recognition method on the CAS–PEAL database.

	Pose	Illumination
Situation estimation	13.6	46.4
Feature extraction	3.7	3.7
Local similarity evaluation and rank list creation	0.5	0.1
Gallery comparison	2.3	0.2
Total	20.1	50.4

each situation of the model set, good matches have been added to the bunch graph to achieve also a good match on previously poor matches. This is repeated until all model images are in the bunch graphs $\mathcal{M}^v$ belonging to their situation. This leads to improved landmark finding (see (Heinrichs et al., 2006) for full details of the method).

For aligning a gallery image, a single match has to be performed with the bunch graph $\mathcal{M}^0$ of the standard situation. After that, similarities to all model images are calculated and the rank lists are created.

Identifying a probe image in situation v works as follows. A single match with the bunch graph $\mathcal{M}^v$ of the appropriate situation has to be done for landmark finding. A comparison with each model subject is done to calculate the rank lists. Then the rank lists can be compared to the ones from the gallery in an all-to-all comparison.

The tests with different combinations of local similarity functions are shown in Table 1. In the following evaluations, the best combination was applied.

Table 2 shows the computation times for the components of the recognition procedure. They are clearly dominated by the situation estimation, which is a crude and inefficient method. Much better ones are available, especially in the case of pose differences (Murphy-Chutorian & Trivedi, 2009). These can be used preceding the rank list evaluation without any change to the application of the learned invariances.

4.3. Robustness of the spiking neural network

We have tested the neural network on the pose and illumination variations of the CAS–PEAL database (Gao et al., 2008, 2004). Landmarks and bunch graphs were created exactly as described in Section 4.2. As similarities we used the combination of $S_{\text{Abs}} \cdot S_{\text{Canb}}$, and S_{LGBPHS}, which yields optimal recognition rates. The firing occurred at times according to Eq. (24).

As was expected, the network achieves precisely the same recognition rates as the rank list comparison. A critical question is if it still does so in the presence of disturbances, which would be inevitable in any real system relying on precise timing.

4.3.1. Random noise

First we have added random offsets $\chi(d)$ both equally or Gaussian distributed with a standard deviation of d to the spike timings of (24) and measured the recognition rate:

$$t_i = 1 - S(J_i^M, J_i^G) + \chi(d). \tag{25}$$

As spike timing is the only carrier of information in the network, it is clear that recognition must decline when noise is added. However, the results in Fig. 7(a) show that small amounts (around 0.05 time units) can still be tolerated.

4.3.2. Early stopping

In an additional experiment on the CAS–PEAL dataset, the decision was made on the basis of subsets of the k most similar model candidates. Neuronally, this means a decision was already made when the first k spikes had reached the gallery neurons. The resulting recognition rates are shown in Fig. 7(b). This shows that recognition rates are not impaired if only the 10–20 most similar model candidates are used. Thus, identity decisions can be made even faster if the gallery neurons do not wait for all spikes to come in.

4.3.3. Dependence on the size of the model gallery

Model learning is only useful if the number of individuals in the model can be much smaller than the number of people in the gallery. We have tested different model sizes with a fixed gallery size of 500 individuals for pose and 91 for illumination. The results are shown in Fig. 7(d). The curves show that the recognition rate for pose saturates around a model database size of 300. This small number of examples seem to suffice to learn the pose transformation for many more people, even of different ethnicity (Table 3). For illumination this saturation was not achieved due to a lack of training examples, but it may be expected around the same number (Fig. 7(d)).

Table 3
Overview of the recognition rates (in %) of all experiments except FRGC.

Test set	Model set				
	CAS–PEAL pose	CAS–PEAL illum.	CAS–PEAL pose + PIE	CAS–PEAL pose + SCface pose	SCface surveillance
CAS–PEAL pose	98.9				
CAS–PEAL illum.		88.9			
PIE	69.1			69.1	
SCface pose	48.8		48.8		
CAS–PEAL + PIE + SCface pose	84.4				
SCface surveillance					20.5

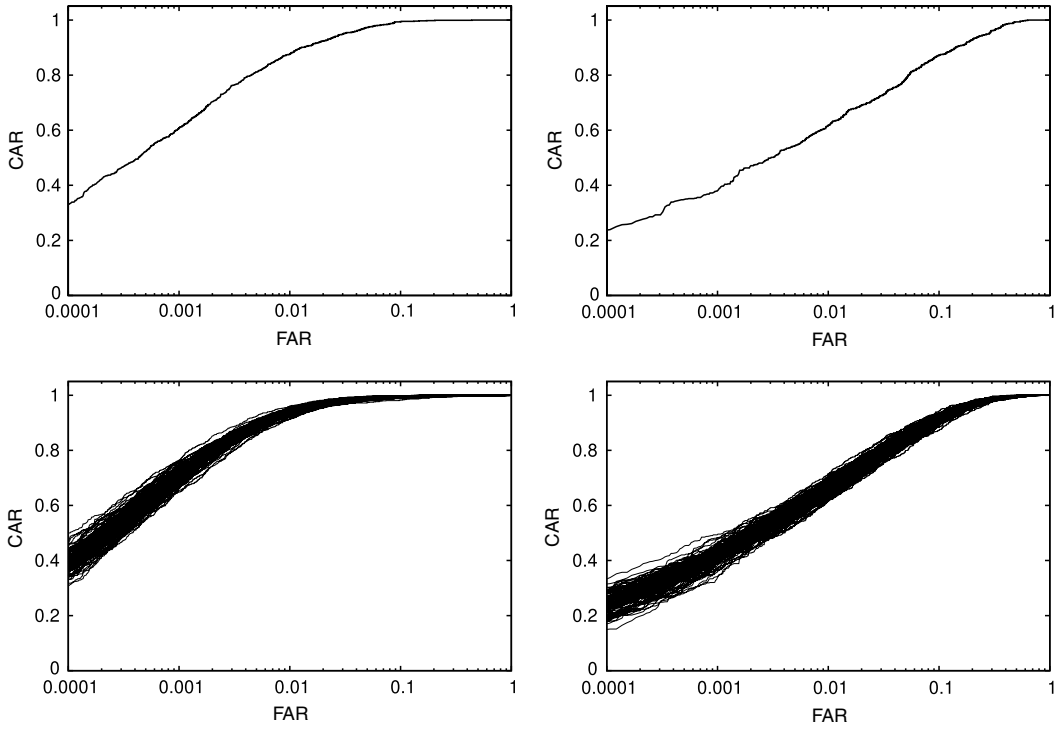

Fig. 6. ROC curves for verification on the CAS–PEAL database. The left column shows the results for pose variation, the right for illumination variation. The lower graphs show the ROC curves for 100 random different partitions of the identities into model and test set. The relatively small variation of the curves shows that the verification capabilities do not depend critically on the selection of models.

4.3.4. Multiple spikes

The assumption that an activated feature detector would fire only a single spike at a precise time is not in accordance with neurophysiology. The general view is that activation causes a spike train, with activity being coded in the *frequency* of spikes (for a recent discussion of evidence see Rolls & Treves, 2011). In a second simulation the active neurons created a volley of spikes, which lasted for $T = 3$ time units:

$$t_i(n) = n \cdot (1 - S(J_i^M, J_i^G)),$$
$$n \in \left\{1, 2, \ldots \frac{T}{1 - S(J_i^M, J_i^G)}\right\}. \tag{26}$$

Subsequent spikes interfere with the evaluation of the rank lists, because they cannot be distinguished from first spikes. The results in Fig. 7(d) show that this does not impede recognition performance.

4.4. Further recognition results

4.4.1. Verification on CAS–PEAL

Beyond the pure recognition rate, the relationship between false positive and false negative decisions in an identity *verification* task is an important measure of the method's quality. In such a scenario the presumed identity is known. The system has to decide if the probe image actually belongs to that identity. As the probe image may be in a different situation, this decision is critical and an important application scenario for situation independence.

We have used all images in situation $v = 0$ of the testing set as gallery images and the ones in different situations as probe. Using the model set they have been compared with all gallery images. If the similarity was above a threshold the identity was accepted, otherwise rejected. This decision is correct if acceptance occurred between images of the same identity or rejection between images of different identities. Varying the threshold yields an ROC for this decision.

Fig. 6 shows this ROC for the described scenario. The correct acceptance rate (CAR) at 0.1% false acceptance rate (FAR) is 60.8% for pose and 37.9% for illumination. The equal error rate (EER) is 4.0% (pose) respectively 11.8% (illumination).

To estimate the dependency of verification on the partitioning of the data, the available subjects have been assigned to model or test in 100 randomly chosen partitions. In the lower half of Fig. 6, we show the collection of all these ROC curves. The relatively small variation shows that the selection of model identities is not crucial for the success of the method or, put more blandly, no matter on which people the change of situation is learned, it can be generalized to the rest.

4.4.2. The PIE database

Using the trained CAS–PEAL model, we reach a recognition rate of 69.1% on the PIE database. Thinking of the difference between model and test set in the form of illumination and race of the subjects, this can be regarded as relatively successful.

4.4.3. The SCface database—pose variation

The recognition rate for the SCface database reaches 48.8%. As an additional difficulty presented by this database, pose angles have a relatively large variation.

4.4.4. Combined databases

Using CAS–PEAL and SCface images as model for the PIE testing set, the recognition rate stays the same. This is also the case in CAS–PEAL and PIE as the model for SCface as testing set. The images of PIE and SCface seem to be too different to be similar at rank list creation and so to influence recognition. The recognition rate in the case of all three databases as test set is 84.4%. See Table 3 for all pairings of model and gallery we have investigated.

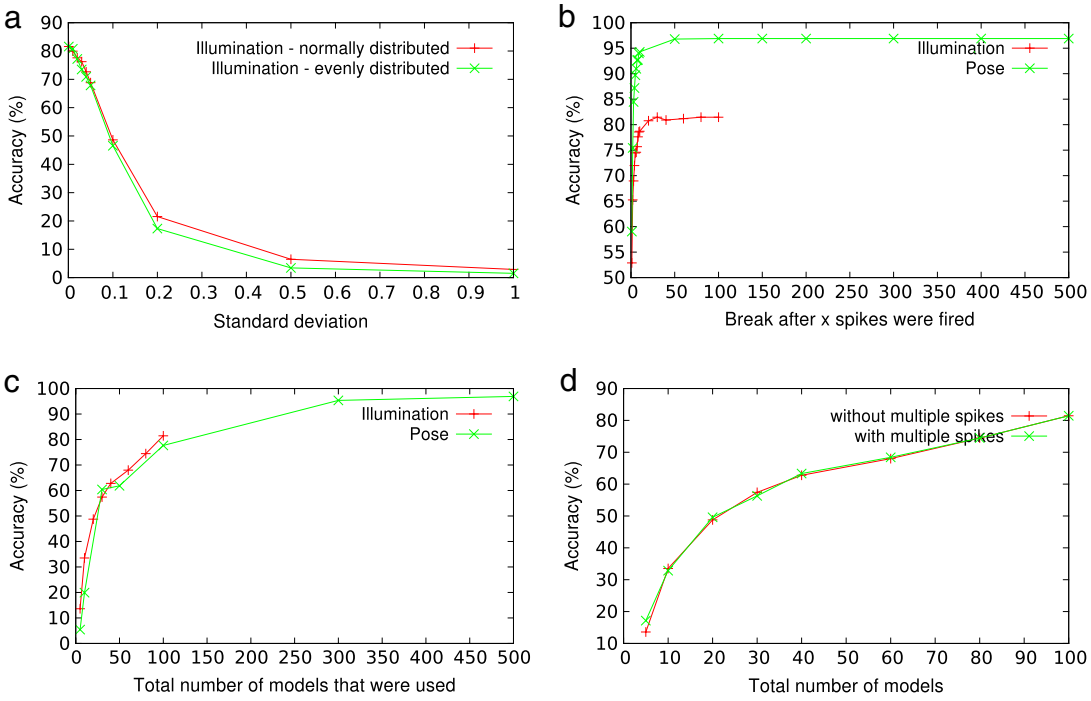

Fig. 7. Results of experiments on the spike-based network. In 7(a) noise was added to the spike times, which are the only carrier of information. 7(b) shows the recognition rate if the spike evaluation is stopped early. 7(c) shows the dependence of the recognition rate on the size of the model database, and 7(d) shows that it does not make a difference if the sent spikes are part of a longer spike train with a frequency coding for the similarity.

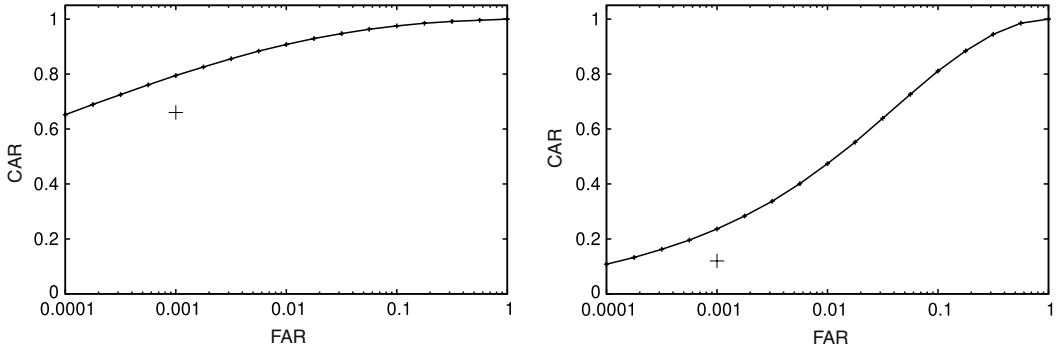

Fig. 8. ROC for FRGC experiment 2.1 with a CAR of 79.5% at an FAR of 0.1% (left) and for FRGC experiment 2.4 with a CAR of 23.6% at an FAR of 0.1% (right). The crosses mark the verification rates of the reference method (PCA).

4.4.5. The SCface database—surveillance camera images

Because of the poor quality of the surveillance camera images this is a very difficult task. With identities 2–81 as model, a recognition rate of 20.5% was achieved. With randomized partitioning into model and test set, the recognition rates were 21.4% ± 2.0%. This shows that the variation introduced by the different cameras could be learned to a certain degree. For comparison, the baseline recognition rates for PCA as reported by Grgic et al. (2009) varied between 0.7% and 8.5%. The authors also discuss that the sample used for training may be just too small to capture the variability in the surveillance images. We would also expect the performance to improve significantly with larger model sets. We found only one other study that does recognition on these images (Choi, Ro, & Plataniotis, 2011). They report recognition rates around 50%, at the cost of manually cropping all images to standard scale and resolution.

4.4.6. The FRGC database

Fig. 8 shows ROCs for the two FRGC experiments. In experiment 2.1 CAR at 0.1% FAR reaches 79.5%, showing an improvement over the reference method (PCA) at 66%. In experiment 2.4, the improvement is even larger at 23.6% compared to 12% for PCA. In this experiment, variation between the images is larger. The improved CAR shows that variations are successfully learned. As part of the FRGC protocol, we compare our results to the baseline, which is defined as PCA. Our results are below the median of all contestants published by Phillips et al. (2006) but close to the mean, which is only published informally by Phillips (2005, p. 67). This is a respectable result given that the contestants are full-blown commercial recognition systems, which invest a lot more effort than our simple learning principle.

5. Discussion

We have presented a face recognition system, which is capable of learning the variations caused by pose and illumination changes and partly from different cameras strictly from examples. The model database holding the variations for a limited number of persons allows the generalization to identities known only in a single situation. The high recognition rates in comparison with previously published recognition results on various databases (see Table 3 for an overview) demonstrate that a usable model of

the variations due to pose and illumination changes has been learned from examples. Results on the FRGC database show that variation can also be learned from a model that is only labeled by identity and not with situation. There are some publications that achieve higher recognition rates on the CAS–PEAL database than our method, (Luo et al., 2007; Rana, Liu, Lazarescu, & Venkatesh, 2008), but they use extensive manual preprocessing for cropping and normalizing the images. We have not found fully automatic recognition methods tested on the CAS–PEAL database.

Clearly, the method must be accompanied by an efficient estimation of situation, which is ongoing research beyond the scope of this paper.

We have also shown that the procedure can be carried out by a neural network based on spike timing in a way robust enough to make it a prime candidate for both biological modeling and massively parallel implementation.

Acknowledgments

We gratefully acknowledge funding from the German Research Foundation (WU 314/2-2 and WU 314/5-2). Portions of the research in this paper use the CAS–PEAL face database collected under the sponsorship of the Chinese National Hi-Tech Program and IS VISION Tech. Co. Ltd. (Gao et al., 2008, 2004). Portions of the research in this paper use the SCface database of facial images (Grgic et al., 2009). Credit is hereby given to the University of Zagreb, Faculty of Electrical Engineering and Computing for providing the database of facial images. Partial results of this paper have been presented at conferences before Günther, Müller, and Würtz (2010), Müller et al. (2007), Müller and Würtz (2009), Müller, Tremer, Bodenstein, and Würtz (2011), and in the theses Bodenstein (2011), Müller (2010) and Tremer (2011).

References

Ayinde, O., & Yang, Y.-H. (2002). Face recognition approach based on rank correlation of Gabor-filtered images. *Pattern Recognition*, *35*(6), 1275–1289.

Bartlett, M. S., & Sejnowski, T. J. (1998). Learning viewpoint-invariant face representations from visual experience in an attractor network. *Network: Computation in Neural Systems*, *9*(3), 399–417.

Bodenstein, C. (2011). Theory and continuous time implementation of a spike time based neural network. B.Sc. Thesis. ET-IT Dept., Univ. of Bochum, Germany.

Choi, J. Y., Ro, Y. M., & Plataniotis, K. N. (2011). A comparative study of preprocessing mismatch effects in color image based face recognition. *Pattern Recognition*, *44*(2), 412–430.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*(2), 194–200.

Franzius, M., Wilbert, N., & Wiskott, L. (2011). Invariant object recognition and pose estimation with slow feature analysis. *Neural Computation*, *23*(9), 2289–2323.

Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., et al. (2008). The CAS–PEAL large-scale Chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, *38*(1), 149–161.

Gao, W., Cao, B., Shan, S., Zhou, D., Zhang, X., & Zhao, D. (2004). The CAS–PEAL large-scale Chinese face database and baseline evaluations. Tech. Rep. JDL-TR-04-FR-001. *Joint research & development laboratory for face recognition*. Chinese Academy of Sciences.

Gonzáles, D., Bicego, M., Tangelder, J. W. H., Schouten, B. A. M., Ambekar, O., Alba-Castro, J. L., et al. (2007). Distance measures for Gabor jets-based face authentication: a comparative evaluation. In S.-W. Lee, & S. Z. Li (Eds.), *LNCS*: *Vol. 4642. Advances in biometrics, ICB 2007* (pp. 474–483). Berlin, Heidelberg: Springer-Verlag.

Grgic, M., Delac, K., & Grgic, S. (2009). SCface—surveillance cameras face database. *Multimedia Tools and Applications*, 1–17.

Günther, M., Müller, M. K., & Würtz, R. P. (2010). Two kinds of statistics for better face recognition. In T. Simos, G. Psihoyios, & C. Tsitouras (Eds.), *Numerical analysis and applied mathematics, international conference* (pp. 1901–1904). American Institute of Physics.

Günther, M., & Würtz, R. P. (2009). Face detection and recognition using maximum likelihood classifiers on Gabor graphs. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(3), 433–461.

Heinrichs, A., Müller, M. K., Tewes, A. H., & Würtz, R. P. (2006). Graphs with principal components of Gabor wavelet features for improved face recognition. In G. Cristóbal, B. Javidi, & S. Vallmitjana (Eds.), *Information optics: 5th international workshop on information optics: WIO'06* (pp. 243–252). American Institute of Physics.

Hinton, G. (1987). Learning translation invariant recognition in massively parallel networks. In G. Goos, & J. Hartmanis (Eds.), *Lecture notes in computer science*: *Vol. 258. PARLE parallel architectures and languages Europe* (pp. 1–13). Springer.

Jitsev, J., & von der Malsburg, C. (2009). Experience-driven formation of parts-based representations in a model of layered visual memory. *Frontiers in Computational Neuroscience*, *3*(15), 1–18.

Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., et al. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, *42*(3), 300–311.

Lücke, J., Keck, C., & von der Malsburg, C. (2008). Rapid convergence to feature layer correspondences. *Neural Computation*, *20*(10), 2441–2463.

Luo, J., Ma, Y., Takikawa, E., Lao, S., Kawade, M., & Lu, B.-L. (2007). Person-specific sift features for face recognition. In *Proc. ICASSP* (pp. II-593–II-596). IEEE.

Ma, B., Zhang, W., Shan, S., Chen, X., & Gao, W. (2006). Robust head pose estimation using LGBP. In *Proc. ICPR. vol. 2* (pp. 512–515).

Müller, M. K. (2010). Lernen von Identitätserkennung unter Bildvariation. Ph.D. Thesis. Physics Dept., Univ. of Bochum, Germany.

Müller, M. K., Heinrichs, A., Tewes, A. H., Schäfer, A., & Würtz, R. P. (2007). Similarity rank correlation for face recognition under unenrolled pose. In S.-W. Lee, & S. Z. Li (Eds.), *LNCS*, *Advances in biometrics* (pp. 67–76). Springer.

Müller, M. K., Tremer, M., Bodenstein, C., & Würtz, R. P. (2011). A spiking neural network for situation-independent face recognition. In Hammer, B., Villmann, T. (Eds.), *Proceedings of New Challenges in Neural Computation*. Frankfurt, August 2011. No. 5/2011 in Machine Learning Reports (pp. 62–69).

Müller, M. K., & Würtz, R. P. (2009). Learning from examples to generalize over pose and illumination. In C. Alippi, M. Polycarpou, C. Panayiotou, & G. Ellinas (Eds.), *LNCS*: *Vol. 5769. Artificial neural networks—ICANN 2009* (pp. 643–652). Springer.

Murphy-Chutorian, E., & Trivedi, M. M. (2009). Head pose estimation in computer vision: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(4), 607–626.

Phillips, J. (2005). Welcome to the third FRGC workshop. URL: http://face.nist.gov/frgc/FRGC_WK3_Brief.pdf.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., & Worek, W. (2006). Preliminary face recognition grand challenge results. In *Proceedings 7th international conference on automatic face and gesture recognition* (pp. 15–24). IEEE Computer Society.

Press, W., Flannery, B., Teukolsky, S., & Vetterling, W. (1988). *Numerical recipes in C—the art of scientific programming*. Cambridge University Press.

Rana, S., Liu, W., Lazarescu, M., & Venkatesh, S. (2008). Recognising faces in unseen modes: a tensor based approach. In *Proc. CVPR* (pp. 3660–3667). IEEE.

Rolls, E. T., & Treves, A. (2011). The neuronal encoding of information in the brain. *Progress in Neurobiology*, *95*(3), 448–490.

Rukhin, A. L., & Osmoukhina, A. (2005). Nonparametric measures of dependence for biometric data studies. *Journal of Statistical Planning and Inference*, *131*, 1–18.

Sim, T., Baker, S., & Bsat, M. (2003). The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(12), 1615–1618.

Tan, X., & Triggs, B. (2010). Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, *19*(6), 1635–1650.

Thorpe, S., Delorme, A., & Van Rullen, R. (2001). Spike-based strategies for rapid processing. *Neural Networks*, *14*(6–7), 715–725.

Tremer, M. (2011). Robustness of a spike time based neural network for rank list evaluation. B.Sc. Thesis. ET-IT Dept., Univ. of Bochum, Germany.

Wiskott, L., Fellous, J.-M., Krüger, N., & von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 775–779.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Computation*, *14*(4), 715–770.

Wiskott, L., & von der Malsburg, C. (1996). Recognizing faces by dynamic link matching. *Neuroimage*, *4*(3), S14–S18.

Wolfrum, P., Wolff, C., Lücke, J., & von der Malsburg, C. (2008). A recurrent dynamic model for correspondence-based face recognition. *Journal of Vision*, *8*(7).

Zhang, W., Shan, S., Gao, W., Chen, X., & Zhang, H. (2005). Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition. In *Proc. ICCV* (pp. 786–791).

Zhang, B., Wang, Z., & Zhong, B. (2008). Kernel learning of histogram of local Gabor phase patterns for face recognition. *EURASIP Journal on Advances in Signal Processing*, 1–8.