Learning from examples to generalize over pose and illumination

Marco K. Müller and Rolf P. Würtz

Institute für Neural Computation, Ruhr-University, 44780 Bochum, Germany
[marco.mueller,rolf.wuertz]@neuroinformatik.rub.de

Abstract. We present a neural system that recognizes faces under strong variations in pose and illumination. The generalization is learnt completely on the basis of examples of a subset of persons (the model database) in frontal and rotated view and under different illuminations. Similarities in identical pose/illumination are calculated by bunch graph matching, identity is coded by similarity rank lists. A neural network based on spike timing decodes these rank lists. We show that identity decisions can be made on the basis of few spikes. Recognition results on a large database of Chinese faces show that the transformations were successfully learnt.

Key words: rank order coding, face recognition, pose invariance, illumination invariance, learning from examples, controlled generalization

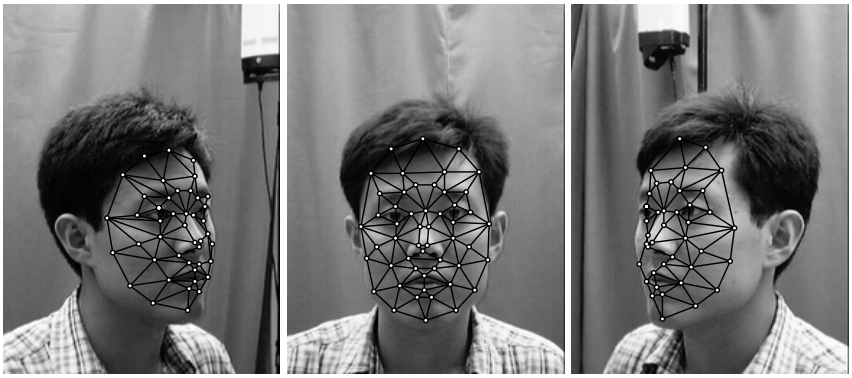
1 Introduction

Invariant recognition of objects is one of the most important features of the visual system and a classical classification task for artificial neural networks. However, invariance is not a natural generalization performed by known network architectures.

Invariances can, to a limited degree, be learnt from real-world data based on the assumption that temporally continuous sequences leave the object identity unchanged [2, 6, 1, 13].

Nevertheless, successful recognition systems have the desired invariances built in by hand. This includes elastic graph matching [7, 12], where the graph dynamics explicitly have to probe all possible variations in order to compare an input image with the stored models. Neural architectures that perform this matching include [14, 8, 15], with the more recent ones being massively parallel and can 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. The latter includes 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 only if bunch graphs are supplied for a coarsely sampled set of variants, e.g., 10 different head poses. This is problematic from a technical point of view [10], because for a



PM+45 PM+00 PM−45

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.

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 supervised way from a set of examples of individual objects in several instances of variations. For lack of a better term, we refer to each coarsely sampled constant illumination or pose angle as one *situation*. Invariant recognition generalizes to other objects that are known only in one situation.

We have recently reported that such a recognition scheme can achieve pose-invariance on the basis of *similarity rank lists* [9]. Here we extend this technique by a neuronal network that implements these similarity rank lists by relative spike timing [11]. This implementation on the one hand gives a plausible neural network for recognition under learnt invariances. On the other hand, it suggests a similarity function, which is different from the one used in [9]. We show that this yields better recognition results for pose-invariant face recognition. In this paper, we also tested the performance on illumination invariance.

2 Recognition by similarity rank lists

Recognition by graph matching [7, 12] compares a given *probe* image P with *gallery images* G_g of all known persons. It first estimates the correspondences between image points on the basis of N local features (Gabor jets) in a process called landmark finding. Then, it calculates a similarity between persons by adding (or averaging) local similarities $S_J(P, G_g, n)$ of *corresponding* features (n being a local feature index). The local similarity function is usually different from the one used for landmark finding. The recognized person is then the G_g

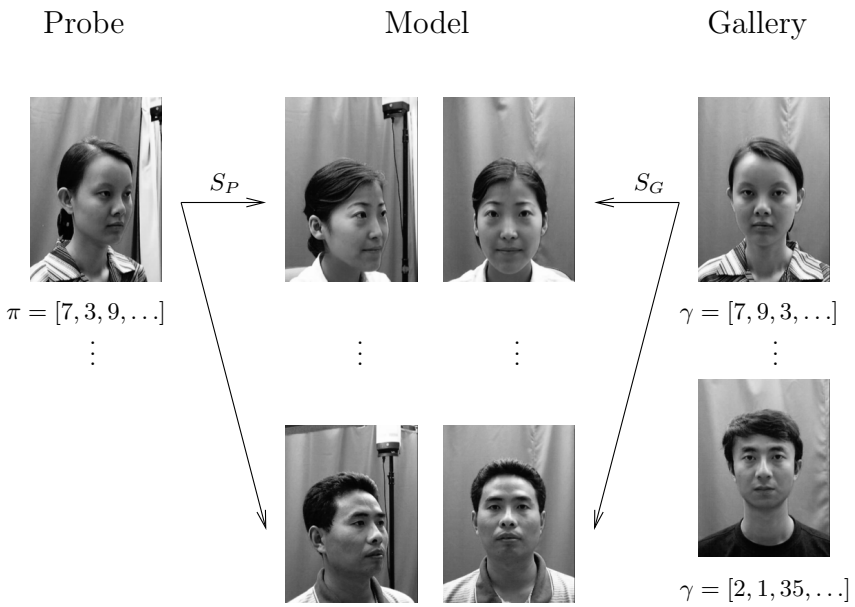


Fig. 2. Situation-independent recognition is mediated by a model database of some persons in all situations. Probe and gallery images are coded into rank lists π and γ by their similarities to the models. These rank lists are comparable, while the similarities are not (feature indices have been dropped for clarity).

with

$$g = \arg\max_g \frac{1}{N} \sum_n S_J(P, G_g, n)\,. \tag{1}$$

This cannot work between different situations, because the features are heavily distorted by pose and illumination changes, for large pose differences some feature points even disappear, leaving no visual features to compare with. In order to overcome this problem we construct a system that can look up the variations in a set of faces which are known in all situations. A number of N_V situations are coded into a *model database* with N_M subjects. The respective graphs are denoted by M_m^v, where m is an index of personal identity and v one of situation. Graphs with the same value of m are derived from images of the same person, the ones with the same value of v show the same situation. On the basis of these examples the variations are learnt.

Each situation requires its own similarity S_v, because the correspondence between features in different situations can not be assumed. Especially, the graphs in different poses contain different numbers of features N^v (see figure 1).

Personal identity is coded by a similarity rank list to the models of the same situation. The rank list for a test subject T 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 r_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 ($r_n^v(m) \in \mathbb{N}_0$):

$$r_n^v(m) < r_n^v(m') \quad \Rightarrow \quad S^v(T, M_m^v, n) \geq S^v(T, M_{m'}^v, n)\,. \tag{2}$$

The most similar model candidate would be the one with $r_n^v(m) = 0$, the follower-up the one with $r_n^v(m) = 1$, etc. These lists now serve as a representation of a test image T. For varying T we will use the notation $r_n^v(T, m)$.

2.1 Invariant recognition

For the recognition of an arbitrary subject a large *gallery* database is created, which contains all known subjects in a preferred situation $v = 0$. For practical purposes, this situation will be a frontal pose under frontal illumination.

Each subject G_g in the gallery is assigned a rank list representation by matching each of its landmarks to those of the model subjects in the preferred situation:

$$\gamma_{g,n}(\cdot) = r_n^0(G_g, \cdot)\,. \tag{3}$$

For recognition we assume that a *probe* P^v image appears in the known situation v. This probe is also represented as a similarity rank list for each landmark of all models in situation v:

$$\pi_n^v(\cdot) = r_n^v(P^v, \cdot)\,. \tag{4}$$

The requirement to know the situation beforehand will be removed in section 2.4.

Now the identity of the probe image is coded into the lists π_n^v, and the gallery images into $\gamma_{g,n}$. Each entry in a rank list is the rank of similarity of that model image to the probe or gallery image.

As the model database contains the same persons in different situations the rank lists should be similar for the same person. This is basically a continuity assumption on the transformations between situations: People that are similar in one situation are also similar in any other situations.

What is required now is a similarity function between rank lists. In contrast to the function chosen in [9] we here construct one on the basis of a neural network, which recognizes patterns on the basis of spike arrival times.

This similarity function enables the comparison of images under pose and illumination variation. For identification tasks it is now sufficient to store a single image of a person in a neutral view. Images taken in different situations can be compared to this gallery image using the rank list similarity.

2.2 Neuronal rank list comparison

Thorpe et al [11] have proposed a neural network that can evaluate rank codes. A set of feature detectors responds to an input pattern such that the most similar

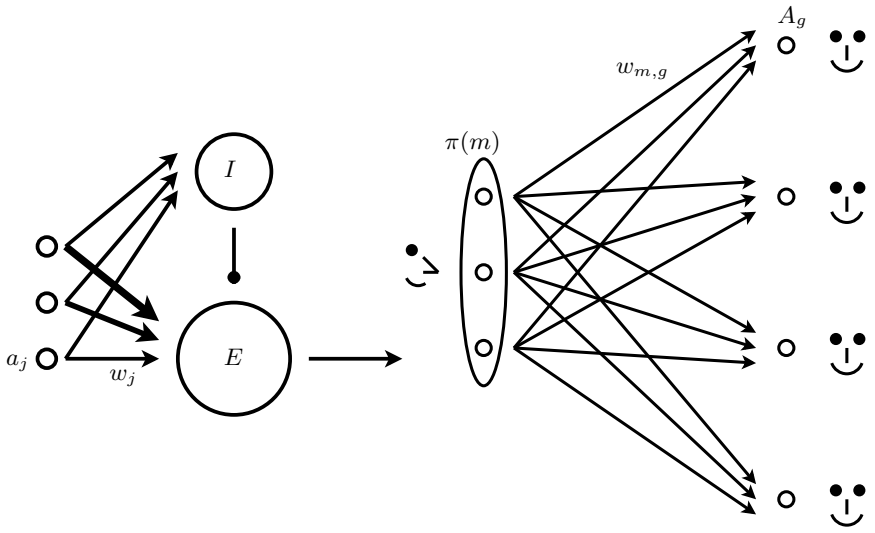


Fig. 3. Left: A neural circuit sensitive to the order of firing neurons, the preferred order is stored in the weights w_j (after [11]). Right; The same circuit is repeated for each gallery image. The probe image is represented as a rank list π according according to similarities with model images in the same situation. The similarities of the gallery to the model images in neutral situation are coded in the weights $w_{m,g}$.

detector fires first. The order in which the spikes arrive can then be decoded by a circuit depicted in the left half of figure 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

$$A = \sum_{j=1}^{K} \exp\left(\frac{\text{order}(a_j)}{\lambda}\right) w_j \,, \tag{5}$$

with λ determining the activity decrease per spike. This parameter has to be optimized, it varies with the size of the rank list. If b_j is the sequence to elicit the largest activation the weights must be

$$w_j = \frac{1}{K} \exp\left(\frac{\text{order}(b_j)}{\lambda}\right) . \tag{6}$$

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

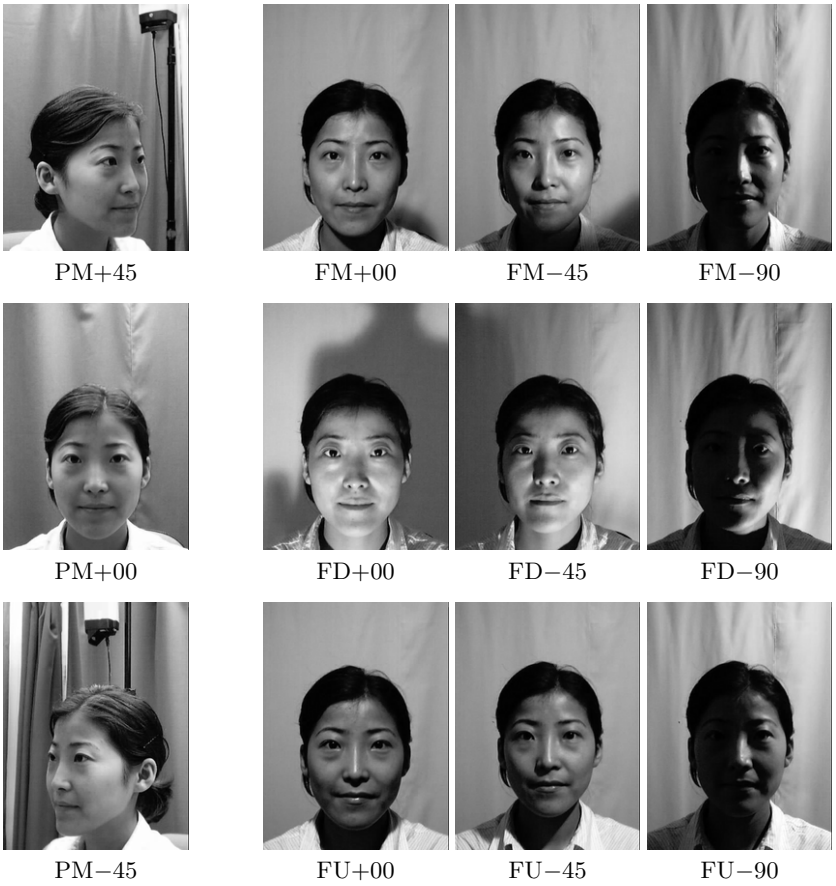


Fig. 4. Examples for pose variation (left column) and illumination variation in frontal pose handled by the system.

$w_{m,g}$ as follows:

$$w_{m,g} = \frac{1}{N_M} \exp\left(\frac{\gamma_g(m)}{\lambda}\right) . \tag{7}$$

The activity A_g then becomes

$$A_g = \sum_m \exp\left(\frac{\pi(m)}{\lambda}\right) w_{m,g} \,, \tag{8}$$

$$= \frac{1}{N_M} \sum_m \exp\left(\frac{\pi(m) + \gamma_g(m)}{\lambda}\right) , \tag{9}$$

and is interpreted as a similarity function between the rank lists π and γ_g.

$$S_{\text{rank}}(\pi, \gamma_g) = \frac{1}{N_M} \sum_{m}^{N_M} \exp\left(\frac{\pi(m) + \gamma_g(m)}{\lambda}\right) . \tag{10}$$

Besides the neural interpretation, this similarity function has yielded better recognition results than the one used in [9].

2.3 Recognition

So far, the feature index n has been omitted from the rank list derivations. Clearly, the above circuit can be repeated for each feature, and the resulting similarities are averaged over all features for a similarity between the persons.

$$S_{\text{recog}}(g) = \frac{1}{N^v} \sum_{n=1}^{N^v} S_{\text{rank}}(\pi_n^v, \gamma_{gn}) . \tag{11}$$

As usual, the recognized person is the one with the index g that maximizes this similarity.

2.4 Automatic estimation of situation

In a realistic setting, the situation of the probe image is, of course, unknown. It can be estimated by matching with bunch graphs of all situations, and assigning the situation with the highest similarity:

$$v_{\text{est}} = \arg\max_v \frac{1}{N^v} \frac{1}{N_M} \sum_{n=1}^{N^v} \sum_{m=1}^{N_M} S^v(T, M_m^v, n) . \tag{12}$$

In case of v situations, bunch graph matching leads to v graphs for a given test image T. 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.

3 Experimental setup

The network was tested on the CAS-PEAL face database [4]. The landmarks are found by elastic bunch graph matching, starting from very few images, that were labeled by hand. 24 subjects have been set aside for manual labeling. From these, the basic bunch graphs have been built (12 for pose, 8 for illumination).

The remaining 1015 subjects have been split up into model sets and testing sets (500 model and 515 testing for the pose case, and 100 model and 91 testing for illumination).

From the basic bunch graphs the landmarks on the model set database have been determined by incremental bunch graph building [9, 5]. After EBGM was

Table 1. Recognition rates (all in %) with known situation are only slightly impaired when the situation is estimated.

	Pose	**Illumination**
Recognition rate with given situation	99.02	89.01
Rate of correct situation estimation	99.89 ± 0.09	91.96 ± 0.89
Recognition rate with automatically determined situation	97.75 ± 0.50	89.97 ± 1.36
Best recognition rate reported in [3]	71	51

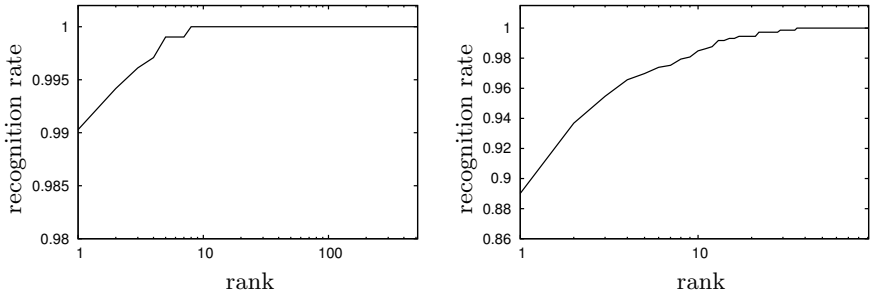


Fig. 5. Cumulative match score with known situation for pose (left) and illumination variation (right). A recognition rate of 100% is reached at rank 8 out of 515 (for pose) and 36 of 91 (for illumination). Rank-1 recognition rates are 99% and 89%, respectively.

performed on one situation of the model set, good matches have been added to the bunch graph to achieve also a good match on previously poor matches. Each situation creates a separate bunch graph. After landmarks for all model images have been found and each bunch graph has grown to a convenient size (15 model graphs have been added in 3 iterations), gallery registration could begin. For registration of a gallery image, a single match has to be performed with the bunch graph of the corresponding situation. After that, similarities to the model images are calculated and the rank lists are created.

Identifying a probe image works as follows. A single match with the bunch graph 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 in the gallery in a cross run.

4 Results

Figure 5 shows the cumulative match scores for recognition under pose and illumination variations. 100% recognition rate has been achieved at rank 8 for pose and 36 for illumination. To estimate the uncertainty in the recognition rate, the available subjects have been assigned to model or test in 100 randomly

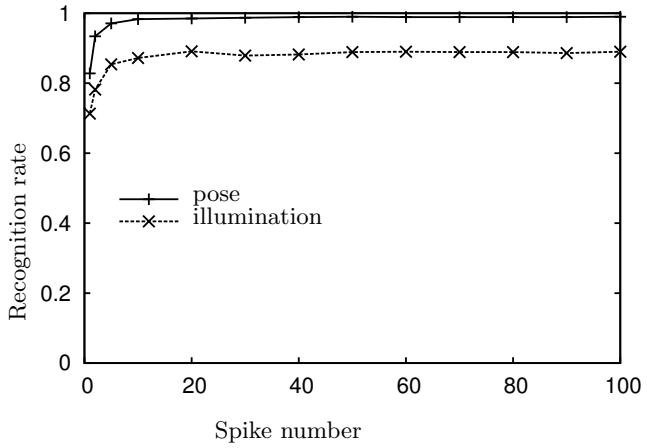


Fig. 6. This curve shows the recognition rates when a recognition decision is made before the spikes from all gallery representations are in. It can be seen that the first 10 spikes suffice to make the correct decision and even the first one is usually a good guess.

chosen partitions. The resulting recognition rates with error bars are shown in table 1.

In a final experiment, the decision was made on the basis of subsets of the k most similar model candidates. This means, a decision was already made when the first k spikes had reached the gallery neurons. The resulting recognition rates are shown in figure 6. This shows that recognition rates are not impaired if only the 10 most similar model candidates are used.

5 Discussion

We have presented a neural network based on spike timing, which is capable of learning the variations caused by pose and illumination changes on the basis of examples. Decisions are made from spike timing with the most similar template firing first. The model database holding the variations for a limited number of persons allows the generalization of identities known only in a single situation. The high recognition rates in comparison with previously published recognition results on the CAS-PEAL database demonstrate that a usable model of the variations due to pose and illumination changes has been learnt from examples. The recognition decision can be made using early stopping, which makes the system very fast in a parallel architecture.

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 ISVISION Tech. Co. Ltd. [4, 3].

References

1. M. S. Bartlett and T. J. Sejnowski. Learning viewpoint-invariant face representations from visual experience in an attractor network. *Network – Computation in Neural Systems*, 9(3):399–417, 1998.
2. P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991.
3. W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao. The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics Part A*, 38(1):149–161, 2008.
4. W. Gao, B. Cao, S. Shan, D. Zhou, X. Zhang, and D. Zhao. The CAS-PEAL large-scale Chinese face database and baseline evaluations. Technical Report JDL-TR-04-FR-001, Joint Research & Development Laboratory for Face Recognition, Chinese Academy of Sciences, 2004.
5. A. Heinrichs, M. K. Müller, A. H. Tewes, and R. P. Würtz. Graphs with principal components of Gabor wavelet features for improved face recognition. In G. Cristóbal, B. Javidi, and S. Vallmitjana, editors, *Information Optics: 5th International Workshop on Information Optics; WIO'06*, pages 243–252. American Institute of Physics, 2006.
6. G. Hinton. Learning translation invariant recognition in massively parallel networks. In G. Goos and J. Hartmanis, editors, *PARLE Parallel Architectures and Languages Europe*, number 258 in Lecture Notes in Computer Science, pages 1–13. Springer, 1987.
7. M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.
8. J. Lücke, C. Keck, and C. von der Malsburg. Rapid convergence to feature layer correspondences. *Neural Computation*, 20(10):2441–2463, 2008.
9. M. K. Müller, A. Heinrichs, A. H. Tewes, A. Schäfer, and R. P. Würtz. Similarity rank correlation for face recognition under unenrolled pose. In S.-W. Lee and S. Z. Li, editors, *Advances in Biometrics*, LNCS, pages 67–76. Springer, 2007.
10. E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.
11. S. Thorpe, A. Delorme, and R. Van Rullen. Spike-based strategies for rapid processing. *Neural Networks*, 14(6-7):715–725, 2001.
12. 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.
13. L. Wiskott and T. J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715 – 770, 2002.
14. L. Wiskott and C. von der Malsburg. Recognizing faces by dynamic link matching. *Neuroimage*, 4(3):S14–S18, 1996.
15. P. Wolfrum, C. Wolff, J. Lücke, and C. von der Malsburg. A recurrent dynamic model for correspondence-based face recognition. *Journal of Vision*, 8(7), 2008.