Pattern Recognition Letters 20(1):89-96 (1999)

The Role of Topographical Constraints in Face Recognition

Laurenz Wiskott *

Institut für Neuroinformatik
Ruhr Universität Bochum, D-44780 Bochum, Germany
and
Computational Neurobiology Laboratory
The Salk Institute for Biological Studies, San Diego, CA 92186-5800, USA

Abstract

The role of topographical constraints for recognition performance is investigated systematically for the case of face recognition. Images are represented by rectangular graphs labeled with jets, based on a Gabor wavelet transform. Topographical constraints are varied between rigid and no constraints. A comparison with two elastic graph matching algorithms is made. The simple methods presented in this paper and elastic graph matching perform comparably on easy galleries, i.e. different facial expression or 11° rotation in depth. On a 22° gallery, elastic graph matching performs significantly better.

Keywords: face recognition, Gabor wavelet transform, topographical constraints.

1 Introduction

One of the main problems in face recognition is to reliably find the face and its landmarks in the first place. In practical systems most of the effort goes into solving this task, while the actual recognition based on features extracted from these facial landmarks is only a minor last step. A typical approach is to match a flexible face model to the image (e.g. Lanitis et al., 1995; Wiskott et al., 1997). The face model can be a graph with nodes encoding local features and edges encoding the geometry of the face. The geometry needs to be flexible to account for individual variations in geometry and distortions due to rotation in depth or mimic expression. It is this flexible geometry of the edges which represents the topographical constraints and which is thought to be crucial for the matching quality and the recognition performance.

Some recent work on object recognition has shown that good recognition results can be obtained without an elaborate matching step and even without any topographical constraints (Mel, 1997; Rao and Ballard, 1995; Viola, 1996). The main objective of this work is therefore to investigate systematically how much topographical constraints contribute to recognition performance in face recognition. In contrast to the systems (Mel, 1997; Rao and Ballard, 1995; Viola, 1996) this work considers different degrees of constraints, ranging from no constraints to very accurate constraints. I refer explicitly to the systems developed by von der Malsburg and collaborators (Lades et al., 1993; Wiskott et al., 1997), for which a direct comparison on the Bochum database is performed. This work also considers some more specific aspects such as the role of phase information in the Gabor jets for matching and recognition.

2 Face Representation

A face is represented by a set of *jets* $\mathcal{J}$ taken from a rectangular grid of pixel positions. This structure is referred to as a *graph*, whose nodes are labeled with jets and edges are labeled with topographical information, i.e. distances or neighborhood relationships. The jets are defined exactly as in (Lades et al., 1993; Wiskott

*Now at the Institute for Advanced Studies, Wallotstr. 19, D-14193 Berlin, Germany, wiskott@wiko-berlin.de

et al., 1997). They are based on a wavelet transform, defined as a convolution of the image with a family of *Gabor kernels* in the shape of plane waves restricted by a Gaussian envelope function. We employ a discrete set of 5 different spatial frequencies and 8 orientations resulting in 40 complex coefficients per pixel. Jets can be compared by two different similarity functions (Wiskott et al., 1997). The first one, $\mathcal{S}_a(\mathcal{J}, \mathcal{J}')$, ignores the phase of the complex coefficients. This is in analogy to using a power spectrum. It yields a smooth potential with large attractor basins if a fixed jet $\mathcal{J}'$ is compared with an array of jets $\mathcal{J}(\vec{x})$ derived from an image. The second one, $\mathcal{S}_\phi(\mathcal{J}, \mathcal{J}')$, takes phase into account. It is more sensitive to displacements and potentially more discriminative since jets with same magnitudes but different phase relations can be distinguished.

3 Database

All experiments are done on the Bochum face database (cf. Lades et al., 1993; Wiskott et al., 1997). For most persons there is one neutral frontal view (fa), one frontal view of different facial expression (fb), and two views rotated in depth by about 11° and 22°, respectively. The neutral frontal views serve as a model gallery, and the other three poses are used as probe images for recognition tests. All images where taken under the same conditions, i.e. illumination and distance from the camera. Thus the faces have a natural variance in size by a factor of up to 1.5. The models, i.e. the neutral frontal views, are represented by 10×10 arrays of jets (see Figure 1). Though the grids are rectangular and regular, i.e. the spacing between the nodes is constant within each dimension, the grids are scaled horizontally and vertically and are aligned manually: The right eye is always represented by the node in the fourth column from the left and the third row from the top, the mouth lies on the fourth row from the bottom, etc. An input image of a face to be recognized is represented by a 16×17 array of jets. The image grids are the same for all images, thus while the model gallery was set up under manual control the recognition of a new face is completely automatic. All galleries used in the experiments below have a size of 108.

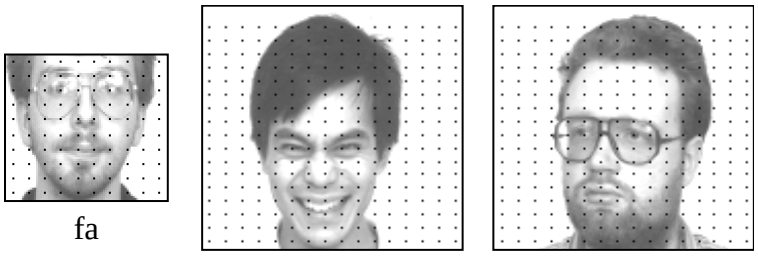

Figure 1: Examples from the Bochum database and the jet locations indicated by black dots.

4 Experiments

Various methods of comparing model graphs G^M with image graphs G^I are considered. The basic similarity function is given by

$$\mathcal{S}\left(G^M, G^I\right) = \max_{d_1 d_2} \frac{1}{N_M} \sum_{m_1 m_2} \max_{p_1 p_2} \mathcal{S}_a\left(\mathcal{J}^M_{m_1 m_2}, \mathcal{J}^I_{(m_1+d_1+p_1)(m_2+d_2+p_2)}\right) . \tag{1}$$

$\mathcal{S}_a(\mathcal{J}^M, \mathcal{J}^I)$ indicates the similarity between a model jet and an image jet, not taking into account phase information. A single model jet $\mathcal{J}^M_{m_1 m_2}$ at node location (m_1, m_2) is usually compared with a square patch of $P \times P$ jets in the image graph. The maximum over all these similarities yields the similarity value $\max_{p_1 p_2} \mathcal{S}_a\left(\mathcal{J}^M_{m_1 m_2}, \mathcal{J}^I_{(m'_1+p_1)(m'_2+p_2)}\right)$ for the model jet, with $p_1 = 0, ..., P-1$ and $p_2 = 0, ..., P-1$. The patches have the same spatial arrangement as the model jets have, i.e. they also form a regular 10×10 grid. This can be achieved by setting $m'_1 = m_1 + d_1$ and $m'_2 = m_2 + d_2$, where (d_1, d_2) indicates the location of the model graph on the image graph. The total similarity of the model graph with the image graph at location (d_1, d_2) is given by the average similarity values of the model jets, i.e.

$\frac{1}{N_M}\sum_{m_1 m_2}\max_{p_1 p_2}\mathcal{S}_a\left(\mathcal{J}^M_{m_1 m_2}, \mathcal{J}^I_{(m_1+d_1+p_1)(m_2+d_2+p_2)}\right)$, where $N_M = 100$ indicates the number of nodes in the model graph. The maximum similarity value over all possible locations (d_1, d_2) is taken as the graph similarity $\mathcal{S}(G^M, G^I)$. The jet correspondences for which this maximum is achieved is the matching result. A face of an image is recognized correctly if the correct model yields the highest graph similarity value.

Topographical constraints can be varied by changing P. $P = 1$ enforces a rigid one to one map without any distortion, though the location (d_1, d_2) is still variable. $P = 6$ represents a weak topographical constraint. In order to test the system without any topographical constraint, comparisons are also made where each node in the model graph is compared with all nodes in the image graph. This is indicated by $P =$ all. Notice that similarity function (1) defines the topographical constraints implicitly by the size of the patches, and it does that on a discrete set of grid points. This is in contrast to the smooth and explicit constraint represented by the second term of Eq. (4) in (Wiskott et al., 1997). I have chosen similarity measure (Eq. 1) rather than Eq. (4) in (Wiskott et al., 1997) mainly for computational convenience, as it permits fast exhaustive search for the global maximum.

Experiment 1: maximum vs. sum scheme. Similarity function (1) is based on the maximum over the similarities of a model jet with the jets of a patch in the image graph. This is not standard in neural models of translation invariant object recognition. The models (Fukushima et al., 1983; Konen et al., 1994; Olshausen et al., 1993), for example, use the (weighted) average instead of the maximum. This may be appropriate for primitive features such as the grey value. For more complex features, however, the maximum scheme turns out to be more efficient than the average scheme. Table 1 shows recognition results for the maximum scheme (Eq. 1) compared with the average scheme for which $\max_{p_1 p_2}$ is replaced by $(1/P^2)\sum_{p_1 p_2}$. The maximum scheme is used in succeeding experiments.

	$P = 1$	2	3	4	5	6	all
average	**72**	**70**	**70**	57	40	26	2
maximum	72	**75**	**76**	70	66	66	33

Table 1: Recognition results in % for the maximum scheme versus the average scheme on the 22° gallery. Matching and recognition was done for individual models without phase. In this and succeeding tables, best recognition results in a row or a column are indicated by boldface and underlined figures, respectively. Differences of up to three percent are considered to be not significant.

Experiment 2: with phase vs. without phase. Similarity function (1) is based on the jet similarity function without phase. It is interesting to see whether using phase can improve recognition performance. It is useful to distinguish between the matching step and the recognition step. In the matching step the correspondences between the nodes are found by similarity function (1), with or without phase, i.e. using jet similarity function $\mathcal{S}_a$ or $\mathcal{S}_\phi$. In the recognition step the similarity of the model to the image is calculated as the average similarity between these corresponding jets, with or without phase. Thus the matching can be done with phase while the recognition is done without phase and *vice versa*. Table 2 shows results for the four different combinations. Matching with phase yields better results than without phase in most of the cases. Recognition without phase has a minor advantage over recognition with phase, an effect hardly noticeable in this experiment but more pronounced if a face bunch graph is used for matching (see below and (Wiskott et al., 1997)). The combination with/without phase is used in succeeding experiments.

matching/recognition	$P = 1$	2	3	4	5	6	all
without/without phase	72	**75**	**76**	70	66	66	33
without/with phase	36	62	**69**	**70**	**69**	63	62
with/without phase	60	74	**79**	**78**	**81**	77	66
with/with phase	56	69	**82**	**79**	78	74	63

Table 2: Recognition results in % for the different combinations of matching with and without phase and recognition with and without phase on the 22° gallery. Matching was done with individual models.

Experiment 3: face bunch graph vs. individual models. In the previous experiments each model

was matched to the image graph independently, i.e. a different set of node correspondences was used for each model. Matching of each individual graph becomes expensive for large galleries but it can be avoided by using the bunch graph technique (Wiskott et al., 1997). A bunch is a set of jets taken from the same node from different model graphs. For example, all jets taken from the node in the third row and fourth column constitute the right eye bunch. This requires, of course, that the model grids are aligned, so that a given node always refers to the same facial location. The most likely right eye node in the image graph is the one whose jet yields the highest similarity value with the right-eye bunch, i.e. with the best fitting jet in the right-eye bunch. The face bunch graph is matched to the image graph once and the established correspondences are used for all models. All 108 models, including the correct one, are used as a face bunch graph here. As an alternative to the face bunch graph one can use an arbitrary model not in the gallery. Results are shown in Table 3. Performance consistently degrades in the order: individual models, face bunch graph, arbitrary model. The difference between individual models and face bunch graph is not significant for the rigid constraint. Thus if tight topographical constraints are used, the face bunch graph is a good alternative to matching each individual model separately. For very different poses, the face bunch graph approach may be superior if a face bunch graph is available for the new pose while the models are given only in the original pose (cf. Wiskott et al., 1997). Figure 2 shows how reliable the matching is in some sample cases. Degradation in matching quality correlates well with degradation of recognition rates (compare Fig. 2 with Tables 3 and 4). However, overall it may be surprising that recognition rates are relatively high given the poor matching quality. For the succeeding experiments, matching individual models is used because it yields the highest recognition rates.

matching with	$P=1$	2	3	4	5	6	all
individual models	60	74	**79**	**78**	**81**	77	66
face bunch graph	59	**68**	**69**	64	55	54	34
arbitrary model	34	**44**	**42**	36	35	28	12

Table 3: Recognition results in % for the different graphs used for matching on the 22° gallery. Matching was done with phase. Recognition was done without phase.

Experiment 4: comparison with elastic graph matching. In the experiments above, topographical constraints were introduced in a rather primitive way, so that it is not surprising that the rigid constraint, $P=1$, yields lower recognition rates than a moderate constraint, e.g. $P=3$. Notice that the model grids scale with the size of the faces while the image grids are kept constant. Thus the grids may be out of scale, which cannot be compensated for under the rigid constraint. Face recognition systems based on more sophisticated matching algorithms (Lades et al., 1993; Wiskott et al., 1997) but using the same jet-representation have been tested on the same database and the identical 108 individuals as used in this paper (see Wiskott et al., 1997). Table 4 summarizes the results. A difference between these three systems consists in the number of jets used: Lades et al. (1993) used 70 jets, Wiskott et al. (1997) used 30 jets, and the system presented here uses 100 jets. However, experiments show that the performance of the latter does not decrease if a sparse subset of 25 jets is used instead of the full set of 100 jets.

pose	L	W	$P=1$	2	3	4	5	6	all
11°	**97**	**94**	85	**94**	92	**94**	**94**	93	91
22°	**85**	**88**	60	74	79	78	81	77	66
fb	**92**	**91**	86	**91**	**92**	**92**	**89**	88	88

Table 4: Recognition results in % in comparison with two versions of a sophisticated matching algorithm (L: Lades et al. (1993); W: Wiskott et al. (1997)). For the simple system presented here, matching was done with phase and individual models. Recognition was done without phase.

System (Lades et al., 1993) matches each individual model separately but uses no phase information. System (Wiskott et al., 1997) uses the face-bunch-graph approach and phase information for matching. Both systems use no phase information for recognition. For the easy galleries, 11° and fb, the simple system

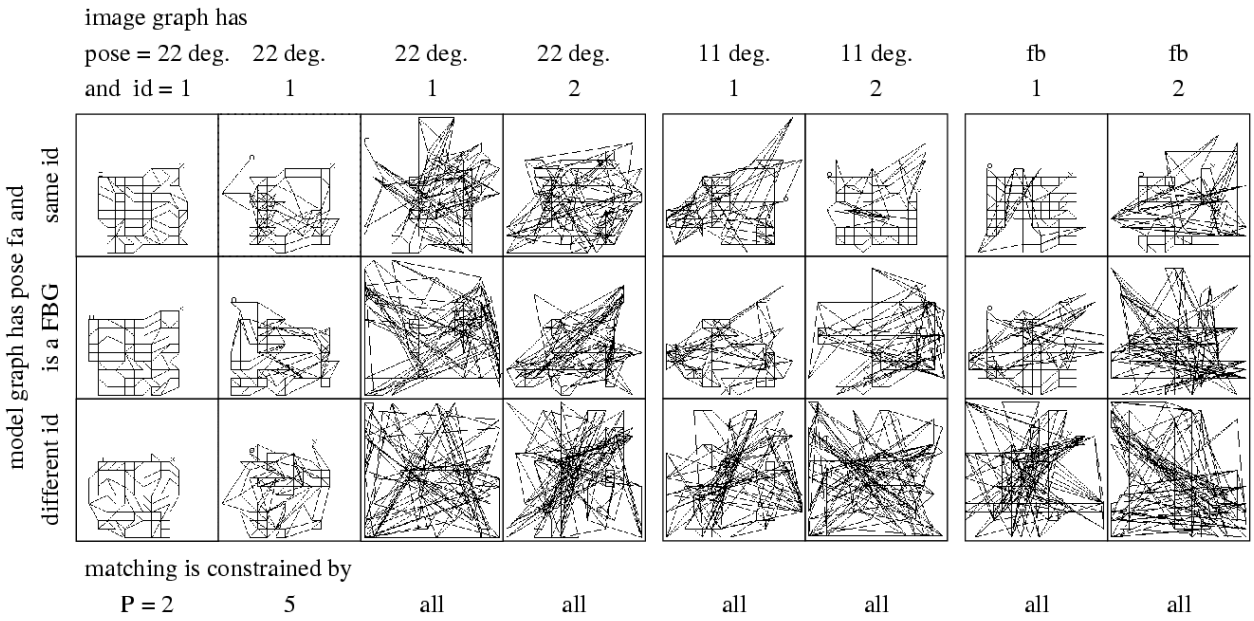

Figure 2: Matching results for various methods. The model graphs are drawn as a 10×10 net in a square representing the image. Model graphs are of pose fa. Image graphs are of pose 22°, 11°, or fb. id 1 and id 2 refer to two different persons. From top to bottom the model graph used for matching is a) the correct model with same identity as the image graph, b) a face bunch graph including the correct model, and c) an arbitrary, i.e. incorrect, model with different identity. The degree to which the matching is topographically constrained is indicated at the bottom. Matching was done with phase.

presented here performs as well as the two more sophisticated algorithms. Even without any topographical constraints there is only little degradation in recognition rates. Thus, for these galleries, topographical constraints seem to play only a minor role. The situation is different for the 22° gallery. Recognition rate without topographical constraints is only two thirds of the optimal performance. However, a simple constraint, such as the one used here with $P = 5$, already improves recognition rates considerably, though a significant performance difference to the elastic graph matching technique remains. It should be noted that the performance of the elastic graph matching algorithms might improve if individual models and phase information would be used for matching, as was was done in the simple system.

5 Discussion

The results of the previous section can be summarized and interpreted as follows:
• Experiment 1 shows that the maximum scheme clearly yields higher recognition rates than the sum scheme. For an interpretation of this result consider the case of $P = 3$ and assume that the model graph is at the correct position on the image graph. Each model jet is compared with a 3×3 patch of image jets of which one is the correct one. One can expect that the correct value is higher than the incorrect ones in most of the cases, since corresponding landmarks should look similar. Therefore, in the maximum scheme the resulting similarity value of the model jet is the correct value with high probability and a higher value otherwise. In the sum scheme the resulting similarity value of the model jet is the average over one correct value and 8 incorrect ones, which leads to a highly unreliable value. This might explain why the maximum scheme yields more reliable similarity values. This consideration only holds for unrelated features, i.e. the nodes need to be far apart. If the nodes within a patch all represent the same texture, for instance, the averaging might actually reduce the noise.
• Experiment 2 shows that the similarity function with phase provides better matching results than the similarity function without phase. There are certain ambiguities that cannot be resolved without phase information. For instance, a light-dark edge yields the same amplitudes as a dark-light edge. With phase,

these ambiguities can be resolved and the matching is more reliable. However, this might hold only if the illumination is fairly constant. With drastically changing illumination, edges can actually change polarity, in which case amplitudes alone might be more robust.

• Experiment 2 also suggests that the similarity function without phase provides slightly better recognition performance, but the results are not strong. It is not clear why the situation is different for recognition than for matching.

• Experiment 3 shows that matching individual models yields higher recognition rates than matching a face bunch graph, which yields higher recognition rates than matching an arbitrary model graph not in the model gallery. The difference between the former two is less pronounced if topographical constraints are imposed. Since matching with an arbitrary model results in poor matching quality (cf. Fig. 2), it is clear that it yields lower recognition rates. The reason for the advantage of matching individual models is not clear, especially since the correct model is also part of the face bunch graph.

• Experiment 4 shows that recognition results are remarkably high for the primitive matching schemes used here. For the easy galleries and if some topographical constraints are imposed the results are comparable to those of sophisticated matching algorithms based on the same feature representation and tested on identical galleries. For the difficult gallery with larger rotation in depth, elastic graph matching performs significantly better than the primitive methods used here. This confirms previous results that reasonable recognition rates can be achieved if the features are complex and the conditions under which the objects have to be recognized do not change too much in terms of rotation in depth and probably also aspects such as illumination, background, scale, etc. Topographical constraints can be useful even if introduced in a fairly crude way.

These results allow interesting conclusions about the potential of different neural systems for translation invariant object recognition. Systems which ignore topographical constraints (e.g. Rosenblatt, 1961) would correspond to column $P =$ all in Table 4. Systems with a fixed routing and some tolerance to local distortions by blurring (e.g. Fukushima et al., 1983; Olshausen et al., 1993) would correspond to columns $P = 1, ..., 6$, depending on the amount of blurring. However, an important difference between the neural systems and the system considered here is that the neural systems usually achieve the distortion tolerance by local averaging while the system here uses the maximum scheme, which is more efficient for complex features. Dynamic link matching (e.g. Bienenstock and von der Malsburg, 1987; Konen et al., 1994; Wiskott and von der Malsburg, 1996) finally corresponds to the columns of systems (Lades et al., 1993; Wiskott et al., 1997). The simple methods perform well if images are similar to the stored model. However, if the generalization over image variation becomes more demanding, the matching methods perform clearly better. A good strategy for biological systems may therefore be to first perform fast recognition with no or little topographical constraints and then refine the recognition process by taking topographical constraints into account.

System (Wiskott and von der Malsburg, 1996) has also been tested on the Bochum gallery. Although it is based on dynamic link matching and should thus perform well, its recognition rates on 111 faces was only 92% on 11°, 66% on 22°, and 85% on fb. This is below the performance of the simple system presented here with $P = 6$ and shows that the system is not well designed. The major flaw is probably that the activity of the blobs instead of the input activity is used for recognition. This may be too sensitive to random fluctuations. An improved version of the model may provide the performance one would expect from the experiments presented here.

This case study has several limitations. Firstly, the background is homogeneous. This would correspond to a situation where an object has been selected by an attention mechanism and segmented by some low level cues. Topographical constraints are probably more important if the image contains a cluttered scene and no selection or segmentation mechanism is applied. Secondly, only faces were considered and the task was an in-class recognition task. It is not clear how the results generalize to different objects. Thirdly, since no ground truth was available as to what the correct matching results would be, some of the experimental results are difficult to interpret, e.g. why is matching individual models more successful than matching a face bunch graph. Further research is needed to clarify these issue. However, the experiments are a first step towards quantifying what the role of topography and various other aspects in object recognition are.

Acknowledgments

I want to thank Christoph von der Malsburg for his support at the Institut für Neuroinformatik and Terrence Sejnowski for his support at the Salk Institute for Biological Studies. This work was supported by a grant from the German Federal Ministry for Science and Technology (413-5839-01 IN 101 B9) and a Feodor-Lynen fellowship by the Alexander von Humboldt-Foundation, Germany.

References

Bienenstock, E. and von der Malsburg, C. (1987). A neural network for invariant pattern recognition. *Europhysics Letters*, 4(1):121–126. 6

Fukushima, K., Miyake, S., and Ito, T. (1983). Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Trans. on Systems, Man, and Cybernetics*, 13:826–834. Also appeared in *Neurocomputing*, J. A. Anderson and E. Rosenfeld, Eds., MIT Press, Massachusetts, pp. 526–534. 3, 6

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

Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., and Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. on Computers*, 42(3):300–311. 1, 2, 4, 6

Lanitis, A., Taylor, C. J., and Cootes, T. F. (1995). An automatic face identification system using flexible appearance models. *Image and Vision Computing*, 13(5):393–401. 1

Mel, B. W. (1997). SEEMORE: Combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation*, 9(4):777–804. 1

Olshausen, B. A., Anderson, C. H., and Essen, D. C. V. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. of Neuroscience*, 13(11):4700–4719. 3, 6

Rao, R. P. N. and Ballard, D. H. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence*, 78:461–505. 1

Rosenblatt, F. (1961). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington, D.C. 6

Viola, P. A. (1996). Complex feature recognition: A Bayesian approach for learning to recognize objects. A.I. Memo 1591, Artificial Intelligence Laboratory of the MIT. 1

Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):775–779. 1, 2, 3, 4, 6

Wiskott, L. and von der Malsburg, C. (1996). Face recognition by dynamic link matching. In Sirosh, J., Miikkulainen, R., and Choe, Y., editors, *Lateral Interactions in the Cortex: Structure and Function*, chapter 11. The UTCS Neural Networks Research Group, Austin, TX, http://www.cs.utexas.edu/users/nn/web-pubs/htmlbook96/. Electronic book, ISBN 0-9647060-0-8. 6