Institut für Neuroinformatik
Ruhr-Universität Bochum

Internal Report 96–05

Face Recognition by Dynamic Link Matching

by

Laurenz Wiskott and Christoph von der Malsburg

Ruhr-Universität Bochum
Institut für Neuroinformatik
44780 Bochum

IR-INI 96–05
March 1996
ISSN 0943-2752

Face Recognition by Dynamic Link Matching*

Laurenz Wiskott†and Christoph von der Malsburg‡
Institut für Neuroinformatik
Ruhr Universität Bochum
D-44780 Bochum, Germany
http://www.neuroinformatik.ruhr-uni-bochum.de

Abstract

We present here a system for invariant and robust recognition of objects from camera images. The system aspires both to be a model for biological object vision (at least an ontogenetically early form of it) and to be at the cutting edge of technological achievement. Our model is based on the principles of temporal feature binding and dynamic link matching. Objects are stored in the form of two-dimensional aspects. These are competitively matched against current images. During the matching process, complete matrices of dynamic links between the image and all models are refined by a process of rapid self-organization, the final state connecting only corresponding points in image and object models. As data format for representing images we use local sets ("jets") of Gabor-based wavelets. We have tested the performance of our system by having it recognize human faces against data bases of more than one hundred images. The system is invariant with respect to retinal position, and it is robust with respect to head rotation, scale, facial deformation and illumination.

The source code for this model is available by anonymous ftp[1] and respective simulation instructions are given in this report.

Keywords: neural networks, dynamic link matching, face recognition, translation invariance, window of attention.

1 Introduction

For the theoretical biologist, the greatest challenge posed by the brain is its tremendous power to generalize from one situation to others. This ability is probably most concretely epitomized in terms of invariant object recognition — the capability of the visual system to pick up the image of an object and recognize that object later in spite of variations in retinal location (as well as other important changes such as size, orientation, changed perspective and background, deformation, illumination and noise). This capability has been demonstrated by flashing the image of novel objects briefly at one foveal position, upon which subjects were able to recognize the objects in a different foveal position (and under rotation in depth) (Biederman & Gerhardstein 1993).

The conceptual grandfather of many of the neural models of invariant object recognition is Rosenblatt's four-layer perceptron (Rosenblatt 1961). It's first layer is the sensory or retinal surface. Its second layer contains detectors of local features (that is, small patterns) in the input layers. Each one of these is characterized by a feature type α and a position x. The third layer contains position-invariant feature detectors, each of which characterized by a feature type α and is to respond to appearance of its feature type anywhere on the input layer. It is enabled to do so by a full set of connections from all of the cells of the

*This work has been funded by grants from the German Federal Ministry of Science and Technology (413-5839-01 IN 101 B/9), from AFOSR (F49620-93-1-0109), from the EU (ERBCHRX–CT–930097), and a grant by the Human Frontier Science Program

†currently at the Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, San Diego, CA 92186-5800, http://www.cnl.salk.edu/CNL, wiskott@salk.edu

‡also Dept. of Computer Science and Section for Neurobiology, University of Southern California, Los Angeles, CA 90089

[1] see Acknowledgement

same feature type in the second layer. Thus, the appearance of a pattern in any position of the input layer leads to the activation of the same set of cells in the third layer. Layer four now contains linear decision units which detect the appearance of certain sets of active cells in the third layer and thus of certain objects imaged into the input layer. A decision unit contains an implicit model of an object in the form of a weighted list of third-layer features to be present or absent.

The four-layer perceptron has to contend with the difficulty that a set of feature types has to be found on the basis of which the presence or absence of a given pattern becomes linearly separable on the basis of the un-ordered feature lists displayed by the third layer. If the feature types employed are too indistinct, there is the danger that different patterns lead to identical third-layer activity, just because the only difference between the patterns is a different spatial arrangement of their features. The danger can be reduced or avoided with the help of feature types of sufficient complexity. However, this is a problematic route itself, since highly complex features are either very numerous (and therefore costly to install) or they are very specific to a given pattern domain (and have to be laboriously trained or hand-designed into the system and limit the system's applicability to the pattern domain). The difficulty arises from the fact, that on the way from layer two to layer three position information is discarded for each feature individually (as is required by the condition of position invariance), such that also information on relative position of the features is lost (which creates the potential confusion).

In the study presented here we are solving the indicated problem using a double strategy. Firstly, we employ highly complex features which are constructed during presentation of individual patterns (and which are stored individually for each pattern later to be recognized), and secondly, we employ a data format and a pattern matching procedure (between our equivalent of Rosenblatt's layers two and three) which represent and preserve relative position information for features.

The features we employ are constructed from image data in a two-step process. First, elementary features in the form of Gabor-based wavelets of a number of scales and a number of orientations are extracted from the image (DAUGMAN 1988), giving a set of response values for each point of the image, then the vector of those response values for a given point are treated as a complex feature, which we call a jet. Jets are extracted from an array of sample points in the image (the approach is described in detail in (LADES et al. 1993)).

Our system is explicit in its representation of analogs for layers two and three, which we call "image domain" and "model domain," respectively. The image domain is an array of (16×17) nodes, each node being labeled by a jet when an image is presented. The model domain is actually a composite of a large number (more than one hundred in some of our simulations) of layers ("models") composed of arrays of (10×10) nodes. To store the image of an object (e.g., a human face) a new model is created in the model domain and its nodes are labeled by copying an array of jets from the appropriate part of the image domain.

To recognize an object, the system attempts to competitively match all stored object models against the jet array in the image domain, a process which we call "Dynamic Link Matching." The winning model is identified as the object recognized. The two domains are coupled by a full matrix of connections between nodes, which is initialized with similarity values between image jets and model jets. (This can be seen as our version of Rosenblatt's feature-preserving connections.) The matching process is formulated in terms of dynamical activity variables for the image and model layers (forming localized blobs of activity in both domains), for the momentary strengths of connections between the domains (we assume that synaptic weights change rapidly and reversibly during the recognition process), and for the relative recognition status of each model. The matching process enforces the condition that neighboring nodes in the image layer link up with neighboring nodes in a model layer. In this way the system suppresses the feature rearrangement ambiguity of the Rosenblatt scheme.

Our model cannot be implemented (at least not in any obvious way) in conventional neural networks. Its implementation is, however, easily possible if two particular features are assumed to be realized in the nervous system, temporal feature binding and rapid reversible synaptic plasticity. Both features have been proposed as fundamental components of neural architecture in (VON DER MALSBURG 1981). Temporal feature binding has in the mean time been widely discussed in the neuroscience literature and has received some experimental basis (KÖNIG & ENGEL 1995). Although rapid synaptic weight changes have been discussed (CRICK 1982) and reported in the literature (ZUCKER 1989), the quasi-Hebbian control and the time course for rapid reversible plasticity that is implied and required here must still wait for experimental validation.

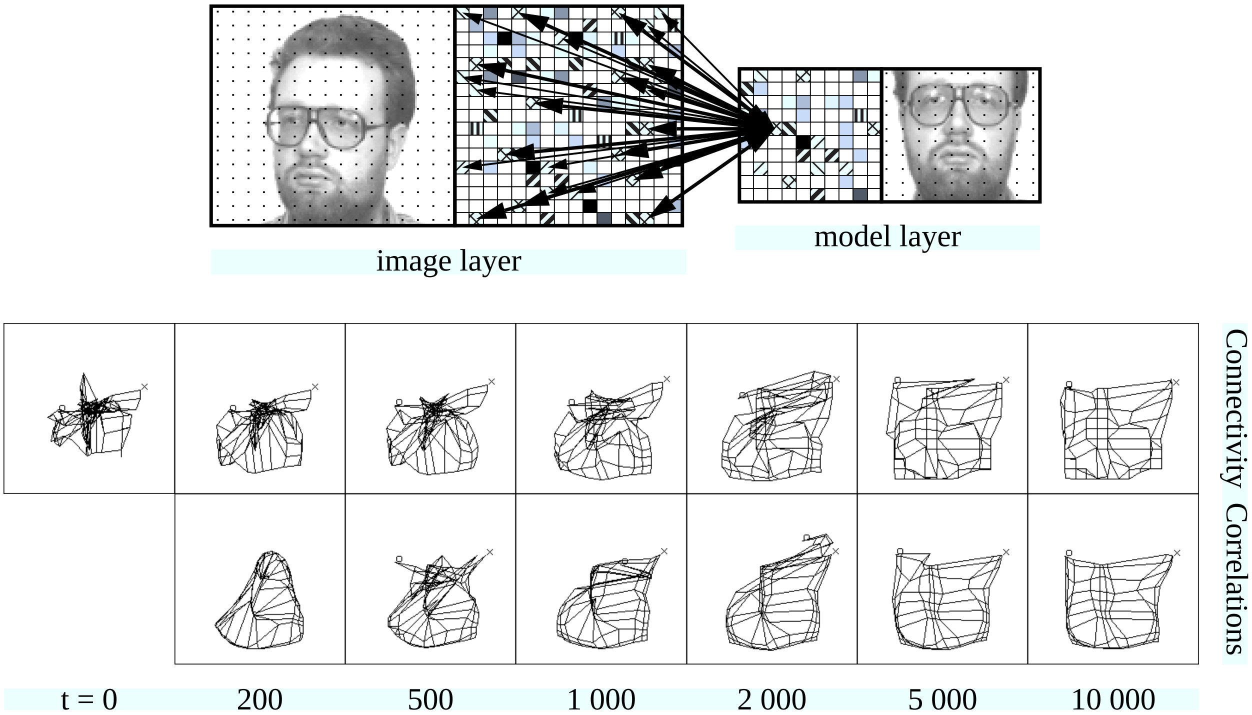

Figure 1: DLM between image and model. The nodes are indicated by black dots, and their local features are symbolized by different textures. The synaptic weights of the initial all-to-all connectivity are indicated by arrows of different line widths. The net displays below show how correlations and connectivity co-develop in time. The image layer serves as a canvas on which the model layer is drawn as a net. Each node corresponds to a model neuron, neighboring neurons are connected by an edge. The nodes are located at the centers of gravity of the projective field of the model neurons, considering synaptic weights as physical mass. In order to favor strong links, the masses are taken to the power of three. The correlations are displayed in the same way, using averaged correlations instead of synaptic weights. It can be seen that the correlations develop faster and are cleaner than the connectivity. The rotation in depth causes a typical distortion pattern; the mapping is stretched on one side and compressed on the other.

2 The System

2.1 Principle of Dynamic Link Matching

In Dynamic Link Matching (DLM), the image and all models are represented by layers of neurons, which are labeled by jets as local features (see Figure 1). Jets are vectors of Gabor wavelet components (see Lades et al. 1993; Wiskott et al. 1995) and a robust description of the local gray value distribution. The initial connectivity is all-to-all with synaptic weights depending on the similarities between the jets. In each layer, neural activity dynamics generates one small moving blob of activity (the blob can be interpreted as covert attention scanning the image or model). If a model is similar in feature distribution to the image, its initial connectivity matrix contains a strong regular component, connecting corresponding points (which by definition have high feature similarity), plus noise in the form of accidental similarities. Hence the blobs in the image and that model tend to align and synchronize in the sense of simultaneously activating, and thus generating correlations, between corresponding regions. These correlations are used, in a process of rapid reversible synaptic plasticity, to restructure the connectivity matrix. The mapping implicit in the signal correlations is more regularly structured than the connectivity itself, and correlation-controlled plasticity thus improves the connectivity matrix. Iteration of this game rapidly leads to a neighborhood preserving one-to-one mapping connecting neurons with similar features, thus providing translation invariance as well as robustness against distortions.

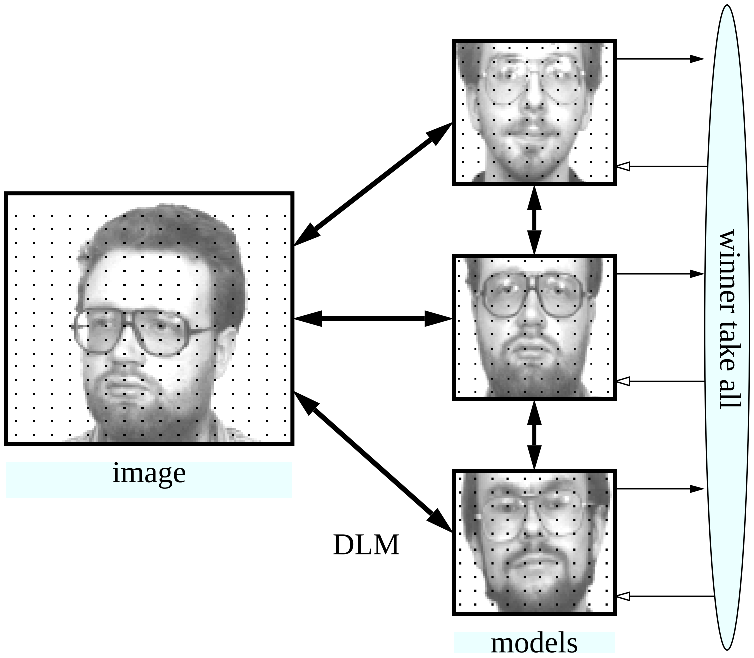

Figure 2: Architecture of the DLM face recognition system. Image and models are represented as neural layers of local features, as indicated by the black dots. DLM establishes a regular one-to-one mapping between the initially all-to-all connected layers, connecting corresponding neurons. Thus, DLM provides translation invariance and robustness against distortion. Once the correct mappings are found, a simple winner-take-all mechanism can detect the model that is most active and most similar to the image.

For recognition purposes, DLM has to be applied in parallel to many models. The best fitting model, i.e. the model most similar to the image, will finally have the strongest connections to the image and will have attracted the greatest share of blob activity. A simple integrating winner-take-all mechanism detects the correct model (see Figure 2).

The equations of the system are given in Table 1; the respective symbols are listed in Table 2. In the following sections, we will explain the system step by step: blob formation, blob mobilization, interaction between two layers, link dynamics, attention dynamics, and recognition dynamics.

2.2 Blob Formation

Blob formation on a layer of neurons can easily be achieved by local excitation and global inhibition (consider Equations 1, 3, and 4 with $\kappa_{hs} = \kappa_{hh} = \kappa_{ha} = \beta_\theta = 0$; cf. also Amari 1977). Local excitation is conveyed by the Gaussian interaction kernel g and generates clusters of activity. Global inhibition, controlled by β_h, lets the clusters compete against each other. The strongest one will finally suppress all others and grow to an equilibrium size determined by the strengths of global inhibition.

Simulation[2]**:** Compile the program with [`prompt> nsl_link DLM.c DLMwin.c`]. In file `DLM.c` the variable `SMALL_LAYER` must be defined, i.e. `#define SMALL_LAYER` instead of `//#define SMALL_LAYER`. Start the simulation with [`prompt> nsl; nsl> load DLMB; nsl> run`], and observe how a blob arises. Restart the simulation with different initial conditions [`Ctrl-C; nsl> init`; mouse clicks with left button on layer `h1`; `nsl> cont`]. Vary also β_h, e.g. [`Ctrl-C; nsl> set data_value beta_h 0.1; nsl> cont`]. What is a reasonable range for β_h?

2.3 Blob Mobilization

Generating a running blob can be achieved by delayed self-inhibition s, which drives the blob away from its current location to a neighboring one, where the blob generates new self-inhibition. This mechanism produces a continuously moving blob (consider Equations 1 and 2 with $\kappa_{hh} = \kappa_{ha} = \beta_\theta = 0$; see also Figure 3). In addition, the self-inhibition serves as a memory and repels the blob from regions recently visited. The driving force and the recollection time as to where the blob has been can be independently controlled by the time constants λ_+ and λ_-, respectively.

[2] see Acknowledgement on how to get the source code

Layer dynamics:

$$
\begin{array}{rcll}
h_i^p(t_0) & = & 0 & \\
\dot{h}_i^p(t) & = & -h_i^p + \sum\limits_{i'} \max\limits_{p'} \left(g_{i-i'}\sigma(h_{i'}^{p'}) \right) - \beta_h \sum\limits_{i'} \sigma(h_{i'}^p) - \kappa_{hs} s_i^p & (1) \\
 & & +\kappa_{hh} \max\limits_{qj} \left(W_{ij}^{pq} \sigma(h_j^q) \right) + \kappa_{ha} \left(\sigma(a_i^p) - \beta_{ac} \right) - \beta_\theta \Theta(r_\theta - r^p) & \\
s_i^p(t_0) & = & 0 & \\
\dot{s}_i^p(t) & = & \lambda_\pm (h_i^p - s_i^p) & (2) \\
g_{i-i'} & = & \exp\left(-\dfrac{(i-i')^2}{2\sigma_g^2} \right) & (3) \\
\sigma(h) & = & \left\{ \begin{array}{rcl} 0 & : & h \leq 0 \\ \sqrt{h/\rho} & : & 0 < h < \rho \\ 1 & : & h \geq \rho \end{array} \right. & (4)
\end{array}
$$

Attention dynamics:

$$
\begin{array}{rcll}
a_i^p(t_0) & = & \alpha_{\mathcal{N}} \mathcal{N}(\mathcal{J}_i^p) & \\
\dot{a}_i^p(t) & = & \lambda_a \left(-a_i^p + \sum\limits_{i'} g_{i-i'} \sigma(a_{i'}^p) - \beta_a \sum\limits_{i'} \sigma(a_{i'}^p) + \kappa_{ah} \sigma(h_i^p) \right) & (5)
\end{array}
$$

Link dynamics:

$$
\begin{array}{rcll}
W_{ij}^{pq}(t_0) & = & \mathcal{S}_{ij}^{pq} = \max\left(\mathcal{S}_\phi(\mathcal{J}_i^p, \mathcal{J}_j^q), \alpha_{\mathcal{S}} \right) & \\
\dot{W}_{ij}^{pq}(t) & = & \lambda_W \left(\sigma(h_i^p)\sigma(h_j^q) - \Theta\left(\max\limits_{j'} (W_{ij'}^{pq} / \mathcal{S}_{ij'}^{pq}) - 1 \right) \right) W_{ij}^{pq} & (6)
\end{array}
$$

Recognition dynamics:

$$
\begin{array}{rcll}
r^p(t_0) & = & 1 & \\
\dot{r}^p(t) & = & \lambda_r r^p \left(F^p - \max\limits_{p'} (r^{p'} F^{p'}) \right) & (7) \\
F^p(t) & = & \sum\limits_i \sigma(h_i^p) &
\end{array}
$$

Table 1: Formulas of the DLM face recognition system

Variables:			
h			internal state of the layer neurons
s			delayed self-inhibition
a			attention
W			synaptic weights between neurons of two layers
r			recognition variable
F			fitness, i.e. total activity of each layer
Indices:			
$(p;p';q;q')$			layer indices, 0 indicates image layer, $1,...,M$ indicate model layers
$=(0;0;1,...,M;1,...,M)$			if formulas describe image layer dynamics
$=(1,...,M;1,...,M;0;0)$			if formulas describe model layers dynamics
$(i;i';j;j')$			two-dimensional indices for the individual neurons in layers $(p;p';q;q')$ respectively
Functions:			
$g_{i-i'}$			Gaussian interaction kernel
$\sigma(h)$			nonlinear squashing function
$\Theta(\cdot)$			Heavyside function
$\mathcal{N}(\mathcal{J})$			norm of feature jet $\mathcal{J}$
$\mathcal{S}_\phi(\mathcal{J},\mathcal{J}')$			similarity between feature jets $\mathcal{J}$ and $\mathcal{J}'$
Parameters:			
β_h	$=$	0.2	strength of global inhibition
β_a	$=$	0.02	strength of global inhibition for attention blob
β_{ac}	$=$	1	strength of global inhibition compensating for the attention blob
β_θ	$=$	∞	global inhibition for model suppression
κ_{hs}	$=$	1	strength of self-inhibition
κ_{hh}	$=$	1.2	strength of interaction between image and model layers
κ_{ha}	$=$	0.7	effect of the attention blob on the running blob
κ_{ah}	$=$	3	effect of the running blob on the attention blob
$\lambda_\pm$			decay constant for delayed self-inhibition
$=\lambda_+$	$=$	0.2	if $h-s>0$
$=\lambda_-$	$=$	0.004	if $h-s\leq 0$
λ_a	$=$	0.3	time constant for the attention dynamics
λ_W	$=$	0.05	time constant for the link dynamics
λ_r	$=$	0.02	time constant for the recognition dynamics
$\alpha_\mathcal{N}$	$=$	0.001	parameter for attention blob initialization
$\alpha_\mathcal{S}$	$=$	0.1	minimal weight
ρ	$=$	2	slope radius of squashing function
σ_g	$=$	1	Gauss width of excitatory interaction kernel
r_θ	$=$	0.5	threshold for model suppression

Table 2: Variables and parameters of the DLM face recognition system

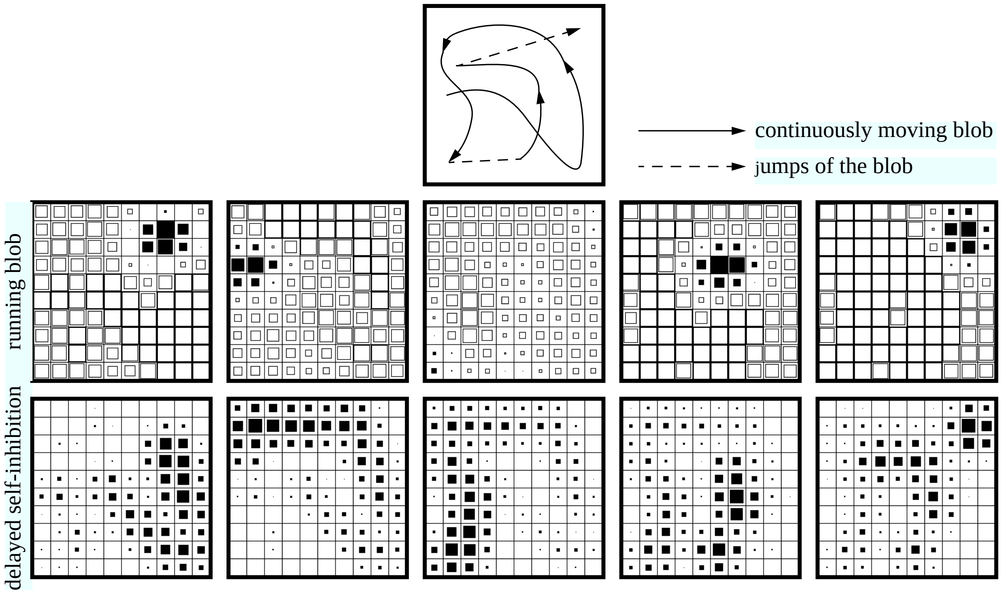

Figure 3: A sequence of layer states. The activity blob h shown in the middle row has a size of approximately six active nodes and moves continuously over the whole layer. Its course is shown in the upper diagram. The delayed self-inhibition s, shown in the bottom row, follows the running blob and drives it forward. One can see the self-inhibitory tail that repels the blob from regions just visited. Sometimes the blob runs into a trap (cf. column three) and has no way to escape from the self-inhibition. It then disappears and reappears again somewhere else on the layer. (The temporal increment between two successive frames is 20 time units.)

Simulation: Start the simulation with [`nsl> load DLMR; nsl> run`], and observe how a blob arises and moves over the layer. Vary λ_+, λ_-, and κ_{hs} (`lambda_p, lambda_m, kappa_hs`), e.g. [`Ctrl-C; nsl> set data_value lambda_m 0.001; nsl> cont`]. Why should λ_- be larger for smaller layers? Is the shape of the blob speed-dependent?

2.4 Layer Interaction and Synchronization

In the same way as the running blob is repelled by its self-inhibitory tail, it can also be attracted by excitatory input from another layer, as conveyed by the connection matrix W (consider Equation 1 with $\kappa_{ha} = \beta_\theta = 0$). Imagine two layers of the same size mutually connected by the identity matrix, i.e. each neuron in one layer is connected only with the one corresponding neuron in the other layer having the same index value. The input then is a copy of the blob of the other layer. This favors alignment between the blobs, because then they can cooperate and stabilize each other. This synchronization principle holds also in the presence of the noisy connection matrices generated by real image data (see Figure 4). (The reason why we use the maximum function instead of the usual sum will be discussed in Section 2.10.)

Simulation: Start the simulation with [`nsl> load DLMS; nsl> run`], and observe how the two blobs synchronize and align with each other. Try different runs (for each run a new object is selected randomly and some synchronize easier than others) and use different object galleries [edit the file `DLMobjects` and exchange the `*pose1` (= 15 degrees rotated faces) block with the`*pose2` (= 30 degrees rotated faces) or `*pose3` (= different facial expression) block]. Vary κ_{hh} (`kappa_hh`). What happens if κ_{hh} is too large or too small?

2.5 Link Dynamics

Links are initialized by the similarity $\mathcal{S}_\phi$ between the jets $\mathcal{J}$ of connected nodes (see Wiskott 1995), with a guaranteed minimal synaptic weight of $\alpha_{\mathcal{S}}$. Then, they become cleaned up and structured on the basis of correlations between pairs of neurons (consider Equation 6; see also Figure 1). The correlations, defined

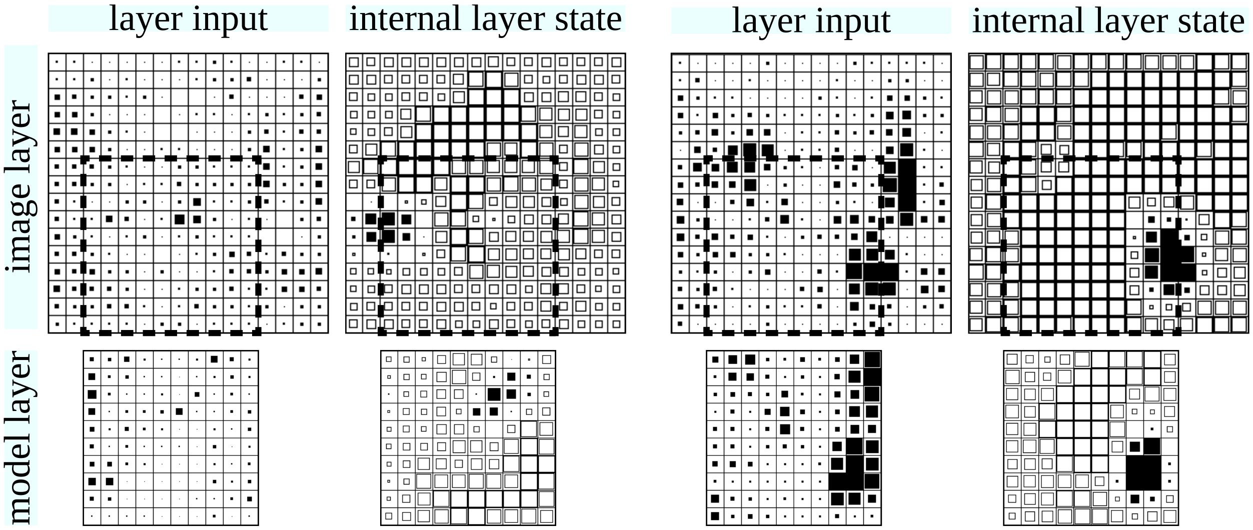

Figure 4: Synchronization between two running blobs. Layer input as well as the internal layer state h is shown at an early stage, in which the blobs of two layers are not yet aligned, left, and at a later state, right, when they are aligned. The two layers are of different size, and the region in Layer 1 which correctly maps to Layer 2 is indicated by a square defined by the dashed line. In the early non-aligned case one can see that the blobs are smaller and not at the location of maximal input. The locations of maximal input indicate where the actual corresponding neurons of the blob of the other layer are. In the aligned case the blobs are larger and at the locations of high layer input.

as $\sigma(h_i^p)\sigma(h_j^q)$, result from the layer synchronization described in the previous section. The link dynamics typically consists of a growth term and a normalization term. The former lets the weights grow according to the correlation between the connected neurons. The latter prevents the links from growing infinitely and induces competition such that only one link per neuron survives, suppressing all others.

Simulation: Start the simulation with [`nsl> load DLMM; nsl> run`], and observe how the connectivity develops in time. Vary λ_W (`lambda_W`). What happens if λ_W is too large?

2.6 Attention Dynamics

The alignment between the running blobs depends very much on the constraints, i.e. on the size and format of the layer on which they are running. This causes a problem, since the image and the models have different sizes. We have therefore introduced an attention blob a which restricts the movement of the running blob on the image layer to a region of about the same size as that of the model layers (consider Equations 1 and 5 with $\beta_\theta = 0$). The basic dynamics of the attention blob is the same as for the running blob, except there is no self-inhibition. Each of the model layers also has the same attention blob to keep the conditions for their running blobs similar to that in the image layer. This is important for the alignment. The attention blob restricts the region for the running blob via the term $\kappa_{ha}\,(\sigma(a_i^p) - \beta_{ac})$, with the excitatory blob $\sigma(a_i^p)$ compensating the constant inhibition β_{ac}. The attention blob on the other hand gets excitatory input $\kappa_{ah}\sigma(h_i^p)$ from the running blob and can thus be shifted into a region where input is especially large and favors activity. The attention blob therefore automatically aligns with the actual face position (see Figure 5). The attention blob layer is initialized with a primitive segmentation cue, in this case the norm of the respective jets (see Wiskott 1995), following the idea that this norm indicates the presence of high contrast texture.

Simulation: Recompile the program with the `SMALL_LAYER` and `SMALL_PATCHES` definitions commented out, e.g. `//#define SMALL_LAYER` instead of `#define SMALL_LAYER`. Start the simulation with [`nsl> load DLMR; nsl> load DLMA; nsl> run`], and observe how an attention blob arises and restricts the region in which the small blob is allowed to move. Vary κ_{ah} and κ_{ha} (`kappa_ah, kappa_ha`). Now restart the simulation with [`nsl> load DLMS; nsl> run`] and see whether the two blobs on the layers of different size can synchronize without an attention blob. Then add the attention blob [`Ctrl-C; nsl> load DLMA; nsl>`

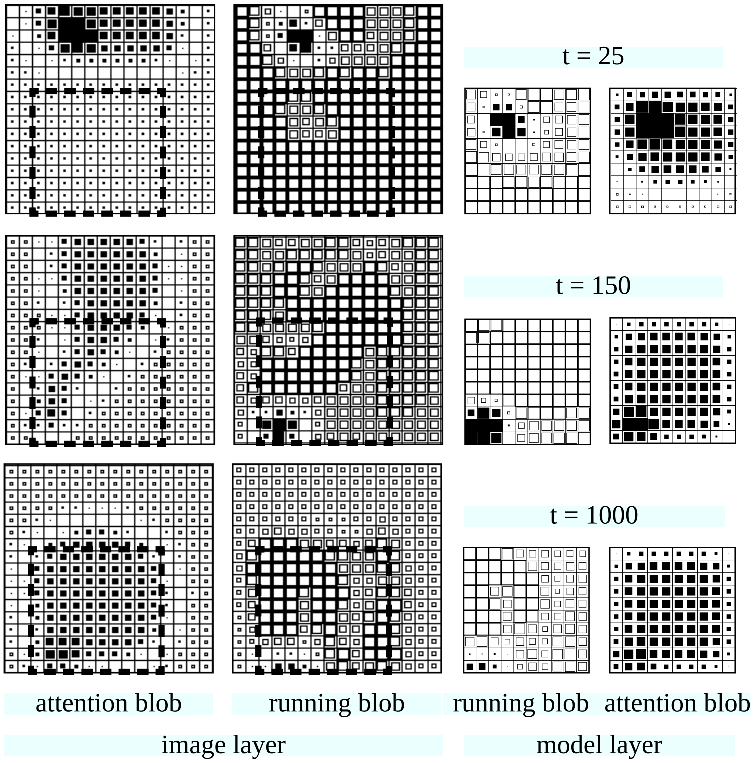

Figure 5: Function of the attention blob, using an extreme example of an initial attention blob manually misplaced for demonstration. At $t = 150$ the two running blobs ran synchronously for a while, and the attention blob has a long tail. The blobs then lost alignment again. From $t = 500$ on, the running blobs remained synchronous, and eventually the attention blob aligned with the correct face position, indicated by a square made of dashed lines. The attention blob moves slowly compared to the small running blob, as it is not driven by self-inhibition. Without an attention blob the two running blobs may synchronize sooner, but the alignment will never become stable.

`run`] and see how the alignment between the blobs can become more stable (notice that for each run a new object is selected randomly, you can suppress that by [`nsl> set data_value ObjectSelectionMode 1`] in which case always the object indicated by `preferredObject` is used; with [`nsl> set data_value ObjectSelectionMode 3`] objects are selected randomly again). You can also experiment with the attention blob misplaced in the beginning [`Ctrl-C; nsl> init`; mouse clicks with the left button near the border on layer `a1`; `nsl> cont`]. Vary again κ_{ah} and κ_{ha}.

2.7 Recognition Dynamics

We have derived a winner-take-all mechanism from Eigen's (1978) evolution equation and applied it to detect the best model and suppress all others (see Equations 1 and 7). Each model cooperates with the image depending on its similarity. The most similar model cooperates most successfully and is the most active one. We consider the total activity of the model layer p as a fitness F^p. The layer with the highest fitness suppresses all others (as can easily be seen if the F^p are assumed to be constant in time and the recognition variables r^p are initialized to 1). When a recognition variable r^p drops below the suppression threshold r_θ, the activity on layer p is suppressed by the term $-\beta_\theta\Theta(r_\theta - r^p)$.

Simulation: Recompile the program with the `SMALL_LAYER` definition commented out but the `SMALL_PATCHES` definition valid, i.e. `//#define SMALL_LAYER` and `#define SMALL_PATCHES`. Start the simulation with [`nsl> load DLMG; nsl> load DLMA; nsl> run`], and observe the recognition process. In the first 1000 time units only the average layer with index 0 is simulated. The correct model has index 1. Shown are, for all models, the total layer activity, the recognition variable and the sum over all synaptic weights (cf. also Figure 6). The connectivity and the layer 2 internal state as well as its input is shown only for the currently most active layer. The time, the index of the most active layer, and the values of the recognition parameters are given as usual output. Asterisks indicate layers which have been ruled out.

2.8 Bidirectional Connections

The connectivity between two layers is bidirectional and not unidirectional as in the previous system (Konen & Vorbrüggen 1993). This is necessary for two reasons: Firstly, by this means the running blobs of the two connected layers can more easily align. With unidirectional connections one blob would systematically run behind the other. Secondly, connections in both directions are necessary for a recognition system. The connections from model to image layer are necessary to allow the models to move the attention blob in the image into a region that fits the models well. The connections from the image to the model layers are necessary to provide a discrimination cue as to which model best fits the image. Otherwise, each model would exhibit the same level of activity.

2.9 Blob Alignment in the Model Domain

Since faces have a common general structure, it is advantageous to align the blobs in the model domain to insure that they are always at the same position in the faces, either all at the left eye or all at the chin etc. This is achieved by connections between the layers, expressed by the term $+\sum_{i'} \max_{p'} \left(g_{i-i'} \sigma(h_{i'}^{p'}) \right)$, instead of $+\sum_{i'} \left(g_{i-i'} \sigma(h_{i'}^{p}) \right)$ in Equation 1. If the model blobs were to run independently, the image layer would get input from all face parts at the same time, and the blob there would have a hard time to align with a model blob, and it would be uncertain whether it would be the correct one. The cooperation between the models and the image would depend more on accidental alignment than on the similarity between the models and the image, and it would then be likely that the wrong model was picked up as the recognition result. One alternative is to let the models inhibit each other such that only one model would have a blob at a time. The models then would share time to match onto the image, and the best fitting one would get most of the time. This would probably be the appropriate setup if the models were of different structure, as is the case for arbitrary objects.

2.10 Maximum Versus Sum Neurons

The model neurons used here use the maximum over all input signals instead of their sum. The reason is that the sum would mix up many different signals, while only one can be correct, i.e. the total input would be the result of one correct signal mixed with many distractions. Hence the signal-to-noise ratio would be low. We have observed an example where even a model identical to the image was not picked as the correct one, because the sum over all the accidental input signals favored a completely different-looking person. For that reason we introduced the maximum input function, which is reasonable since the correct signal is likely to be the strongest one. The maximum rule has the additional advantage that the dynamic range of the input into a single cell does not vary much when the connectivity develops, whereas the signal sum would decrease significantly during synaptic re-organization and let the blobs loose their alignment.

3 Experiments

3.1 Data Base

As a face data base we used galleries of 111 different persons. For most persons there is one neutral frontal view, one frontal view of different facial expression, and two views rotated in depth by 15 and 30 degrees respectively. The neutral frontal views serve as model gallery, and the other three are used as test images for recognition. The models, i.e. the neutral frontal views, are represented by layers of size 10×10 (see Figure 2). Though the grids are rectangular and regular, i.e. the spacing between the nodes is constant within each dimension, the graphs are scaled horizontally and vertically and are aligned manually: The left 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. The x- (that is, horizontal) spacing ranges from 6.6 to 9.3 pixels with a mean value of 8.2 and a standard deviation of 0.5. The y-spacing ranges from 5.5 to 8.8 pixels with a mean value of 7.3 and a standard deviation of 0.6. An input image of a face to be recognized is represented by a 16×17 layer with an x-spacing of 8 pixels and a y-spacing of 7 pixels. The image graphs are not aligned, since that would already require recognition. The variations of up to a factor of 1.5 in the x- and y-spacings must be compensated for by the DLM process.

3.2 Technical Aspects

DLM in the form presented here is computationally expensive. We have performed single recognition tasks with the complete system, but for the experiments referred to in Table 3 we have modified the system in several respects to achieve a reasonable speed. We split up the simulation into two phases. The only purpose of the first phase is to let the attention blob become aligned with the face in the input image. No modification of the connectivity was applied in this phase, and only one average model was simulated. Its connectivity was derived by taking the maximum synaptic weight over all real models for each link: $W_{ij}^{a}(t_0) = \max_{pq} W_{ij}^{pq}(t_0)$. This attention period takes 1000 time steps. Then the complete system, including the attention blob, is simulated, and the individual connection matrices are subjected to DLM. Neurons in the model layers are not connected to all neurons in the image layer, but only to an 8×8 patch. These patches are evenly distributed over the image layer with the same spatial arrangement as the model neurons themselves. This still preserves full translation invariance. Full rotation invariance is lost, but the jets used are not rotation invariant anyway. The link dynamics is not simulated at each time step, but only after 200 simulation steps or 100 time units. During this time a running blob moves about once over all of its layer, and the correlation is integrated continuously. The simulation of the link dynamics is then based on these integrated correlations, and since the blobs have moved over all of the layers, all synaptic weights are modified. For further increase in speed, models which are ruled out by the winner-take-all mechanism are no longer simulated; they are just set to zero and ignored from then on ($\beta_\theta = \infty$). The CPU time needed for the recognition of one face against a gallery of 111 models is approximately 10–15 minutes on a Sun SPARCstation 10-512 with a 50 MHz processor.

In order to avoid border effects, the image layer has a frame with a width of 2 neurons without any features or connections to the model layers. The additional frame of neurons helps the attention blob to move to the border of the image layer. Otherwise, it would have a tendency to stay in the center.

3.3 Results

Figure 6 shows a sample recognition process using a test face strongly differing in expression from the model. The gallery contains five models. Due to the tight connections between the models, the layer activities show the same variations and differ only little in intensity. This small difference is averaged over time and amplified by the recognition dynamics that rules out one model after the other until the correct one survives. The example was monitored for 2000 units of simulation time. An attention phase of 1000 time units had been applied before, but is not shown here. We selected a sample run which had exceptional difficulty to decide between models. The sum over the links of the connectivity matrices was even higher for the fourth model than for the correct one. This is a case where the DLM is actually required to stabilize the running blob alignment and recognize the correct model. In some other cases the correct face can be recognized without modifying the connectivity matrix.

Recognition rates for galleries of 20, 50, and 111 models are given in Table 3. As is already known from previous work (Lades et al. 1993), recognition of depth-rotated faces is in general less reliable than, for instance, recognition of faces with an altered expression. It is interesting to consider recognition times (measured in arbitrary units). Although they vary significantly, a general tendency is noticeable: Firstly, more difficult tasks take more time, i.e. recognition time is correlated with error rate. This is also known from psychophysical experiments (see for example Bruce et al. 1987; Kalocsai et al. 1994). Secondly, incorrect recognition takes much more time than correct recognition. Recognition time does not depend very much on the size of the gallery.

4 Discussion

The model presented here deviates in some very fundamental ways from other biological and neural models of vision or of the brain. Foremost among these is its extensive exploitation of rapid reversible synaptic plasticity and temporal feature binding. Since these features, although first presented a decade and a half ago (von der Malsburg 1981), have not received wide acceptance in the community yet, we have expended great effort to demonstrate the functional superiority of the dynamic link architecture over more conventional

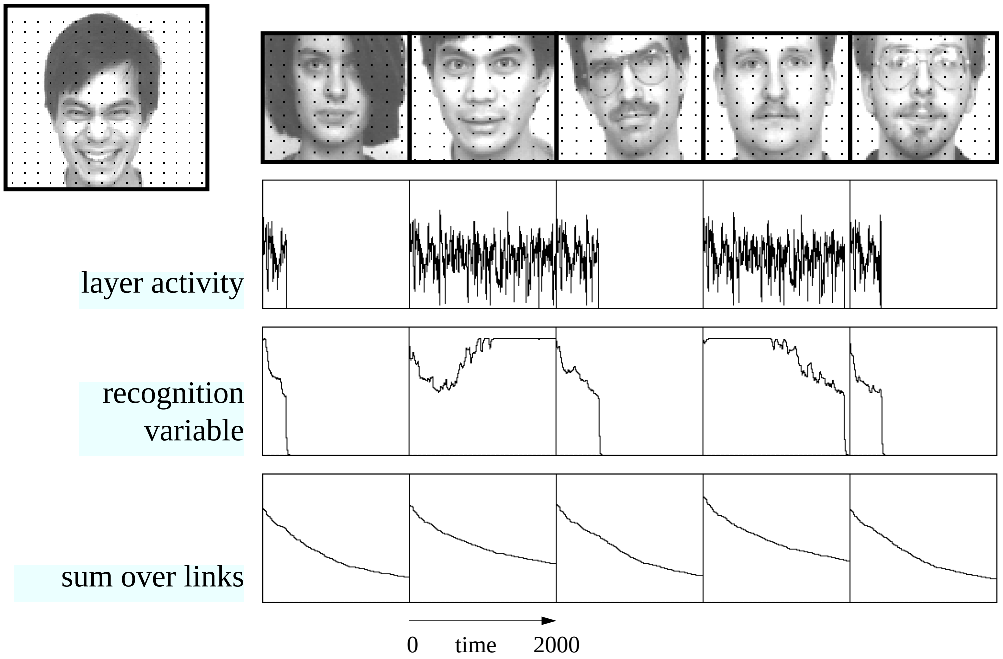

Figure 6: DLM recognition: A sample run. The test image is shown on the left, with 16×17 neurons indicated as black dots. The models have 10×10 neurons and are aligned with each other. The corresponding total layer activities, i.e. the sum over all neurons of one model, are shown in the upper graph. The most similar model is usually slightly more active than the others. On that basis the models compete against each other, and eventually the correct one survives, as indicated by the recognition variable. The sum over all links of each connection matrix is shown in the lower graphs. It gives an impression of the extent to which the matrices self-organize before the recognition decision is made.

Gallery		Correct		Recognition Time for	
Size	Test Images	Recognition		Correct	Incorrect
		#	Rate %	Recognition	Recognition
20	111 rotated faces (15 degrees)	106	95.5	310 ± 400	5120 ±3570
	110 rotated faces (30 degrees)	91	82.7	950 ±1970	4070 ±4810
	109 frontal views (grimace)	102	93.6	310 ± 420	4870 ±6010
50	111 rotated faces (15 degrees)	104	93.7	370 ± 450	8530 ±5800
	110 rotated faces (30 degrees)	83	75.5	820 ± 740	5410 ±7270
	109 frontal views (grimace)	95	87.2	440 ±1000	2670 ±1660
111	111 rotated faces (15 degrees)	102	91.9	450 ± 590	2540 ±2000
	110 rotated faces (30 degrees)	73	66.4	1180 ±1430	4400 ±4820
	109 frontal views (grimace)	93	85.3	480 ± 720	3440 ±2830

Table 3: Recognition results against a gallery of 20, 50, and 111 neutral frontal views. Recognition time (with two iterations of the differential equations per time unit) is the time required until all but one models are ruled out by the winner-take-all mechanism.

neural models by using it to solve a real-world problem, object recognition. We are presenting here our best achievement so far in this venture.

The model presented here is closely related to a more technically oriented system (the "algorithmic system" in contrast to the "dynamical system" described here). It has also been developed in our group and is described in (LADES et al. 1993; WISKOTT et al. 1995). Essential features are common to the two systems, among them the use of jets composed of Gabor-based wavelet features, and of dynamic links to establish a mapping between the image domain and individual models.

Our model for object recognition is successful in emulating the performance and operational characteristics of our visual system in some important aspects. As in the biological case, the flexible recognition of new objects can be installed simply by showing them once. Our system works with a type of standard feature detector, wavelets, which dominates much of the early visual cortical areae (JONES & PALMER 1987). The sensitivity of our system to changes in the stimulus, as for instance head rotation and change in facial expression, is strongly correlated with that of human subjects (KALOCSAI et al. 1994; this study involved a version of our algorithmic system). And, above all, our model is superior in its object discrimination ability to all biologically motivated models known to us, and is at least one of the top competitors among technical systems for face recognition (in a blind test of face recognition against large galleries, performed by the American Army Research Lab, our algorithmic system came out as one of the top competitors, if not the top competitor). Moreover, our system goes beyond mere recognition of objects, providing the basis for a detailed back-labeling of the image with interpretations in terms of explicit object or pattern models which are linked to the image by dynamic links and temporal feature binding.

In spite of this success, there are still some difficulties and discrepancies. One concern is processing time. The reorganization of the connectivity matrix between the image domain and the model domain requires that the two domains be covered at least twice by the running blob. The speed of this blob is limited by the time taken by signal transmission between the domains and by the temporal resolution with which signal coincidence can be evaluated by dendritic membranes and rapidly plastic synapses. Assuming a characteristic time of a few milliseconds we estimate that our model would need at least one second to create a synaptic mapping. This is much too long compared to the adult's speed of pattern recognition (SUBRAMANIAM et al. 1995). We therefore see our system as a model for processes that require the establishment of mappings between the image and object models. This is often the case whenever the absolute or relative placement of parts within a figure is important, and is very likely to be also required when a model for a new object is to be laid down in memory. The actual inspection times required by subjects in such cases are much longer than those required for mere object recognition and can easily be accommodated by our model. We believe that mere recognition can be speeded up by short-cuts. Potential for this we see in two directions, a reduction of the ambiguity of spatial feature arrangement with the help of trained combination-coding features, and a more efficient way (than our running activity blobs) of installing topographically structured synaptic mappings between the image domain and the model domain. A possible scheme for this would be the switching of whole arrays of synapses with the help of specialized control neurons and presynaptic terminals (ANDERSON & VAN ESSEN 1987).

Another as yet weak point of our model is the internal organization of the model domain and the still semi-manual mode in which models are laid down. It is unrealistic to assume completely disjoint models, for several reasons, not the least of which economy in terms of numbers of neurons required. Also, it is unrealistic to see the recognition process as a competition between the dozens of thousands of objects that an adult human may be able to distinguish. Rather, pattern similarities within large object classes should be exploited to give the recognition process hierarchical structure and to support generalization to new objects with familiar traits. The existence of such hierarchies is well supported by neurological observations (DAMASIO & DAMASIO 1992) and is implicit in psychophysical results (BIEDERMAN 1987) showing that many objects are recognized as simple arrays of shape primitives which are universally applicable. In a system closely related to the one presented here (VON DER MALSBURG & REISER 1995), a model domain was dynamically constructed as one comprehensive fusion graph containing as sub-graphs models for different objects, and in fact for different aspects of these objects, with different models sharing many nodes. Further research is required in this direction.

Another limitation of the present system is its inability to deal with alterations of size and orientation of the object image beyond a few percent and beyond a few degrees. For this it would be necessary that the connections between the image domain and the model domain linked also features of different size and orientation. Size and orientation invariance has been successfully demonstrated in the context of the algorithmic system (Buhmann et al. 1990; Lades 1995). Direct implementation in the present model would, however, make the DLM process slower and much more difficult or perhaps even impossible, because the system would have to start with a connectivity matrix with many more non-zero entries. The problem may have to be solved with the help of a two-step DLM process, the first step installing an expectation as to size and orientation of the image, specializing the dynamic links accordingly, the second step organizing the match as described here. In many cases, estimates of size and orientation of an object's image can be derived from available cues, one of which being the object's outline as found by a segmentation mechanism.

In the set of simulations presented here we simplified the recognition problem by presenting the objects to be recognized against a homogeneous background. More difficult scenes may require separate segmentation mechanisms that first identify an image region or regions as candidates for recognition (although a version of the algorithmic system was able to recognize known objects in spite of a dense background of other objects and of partial occlusion (Wiskott & von der Malsburg 1993)). Our model is ideally suited to implement image segmentation mechanisms based on temporal feature binding, as proposed in (von der Malsburg 1981), implemented in (von der Malsburg & Buhmann 1992; Vorbrüggen 1995) and supported by experimental data as reviewed in (König & Engel 1995). According to that idea, all neurons activated by a given object synchronize their temporally structured signals to express the fact that they are part of one segment. This coherent signal, suitably identified with our attention variable a_i^p, Equation 5, could focus the recognition process on segments.

In summary, we feel that in spite of some remaining difficulties and discrepancies we may have, with our model, a foot in the door to understanding important functional aspects of the human visual system.

Acknowledgement

The model has been simulated with the C++ class library NSL (Neural simulation Language) developed by Alfredo Weizenfeld et al. The source code for the model is available by anonymous ftp from `ftp://ftp.neuroinformatik.ruhr-uni-bochum.de/pub/manuscripts/IRINI/irini96-05/irini96-05-src.tar.gz`. Information on NSL, for instance how to get it, can be obtained via WWW from `http://www-hbp.usc.edu:8376/HBP/models/index.html` or via email from Webmaster `<hbp@rana.usc.edu>`.

References

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87.

Anderson, C. H. and Essen, D. C. V. (1987). Shifter circuits: A computational strategy for dynamic aspects of visual processing. *Proc Natl. Acad. Sci. USA*, 84:6297–6301.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147. basic level object classifications can be made in 100 msec.

Biederman, I. and Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *J. Exp. Psychology*, 19:1162–1182.

Bruce, V., Valentine, T., and Baddeley, A. (1987). The basis of the 3/4 view advantage in face recognition. *Applied Cognitive Psychology*, 1:109–120.

Buhmann, J., Lades, M., and von der Malsburg, C. (1990). Size and distortion invariant object recognition by hierarchical graph matching. In *Proceedings of the IJCNN International Joint Conference on Neural Networks*, pages II 411–416, San Diego. IEEE.

Crick, F. (1982). Do dendritic spines twitch? *Trends in Neurobiology*, February:44–46.

Damasio, A. R. and Damasio, H. (1992). Cortical systems underlying knowledge retrieval: Evidence from human lesion studies. In *Neurobiology of Neocortex*. John Wiley.

DAUGMAN, J. G. (1988). Complete discrete 2-d gabor transform by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(7):1169–1179.

EIGEN, M. (1978). The hypercycle. *Naturwissenschaften*, 65:7–41.

JONES, J. AND PALMER, L. (1987). An evaluation of the two dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. of Neurophysiology*, 58:1233–1258.

KALOCSAI, P., BIEDERMAN, I., AND COOPER, E. E. (1994). To what extent can the recognition of unfamiliar faces be accounted for by a representation of the direct output of simple cells. In *Proceedings of the Association for Research in Vision and Ophtalmology, ARVO*, Sarasota, Florida.

KONEN, W. AND VORBRÜGGEN, J. C. (1993). Applying dynamic link matching to object recognition in real world images. In GIELEN, S. AND KAPPEN, B., editors, *Proceedings of the International Conference on Artificial Neural Networks, ICANN*, pages 982–985, London. Springer-Verlag.

KÖNIG, P. AND ENGEL, A. K. (1995). Correlated firing in sensory-motor systems. *Current Opinion in Neurobiology*, 5:511–519.

LADES, M. (1995). *Invariant Object Recognition with Dynamical Links, Robust to Variations in Illumination.* PhD thesis, Fakultät für Physik und Astronomie, Ruhr-Universität Bochum, D-44780 Bochum.

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 Transactions on Computers*, 42(3):300–311.

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

SUBRAMANIAM, S., BIEDERMAN, I., KALOCSAI, P., AND MADIGAN, S. R. (1995). Accurate identification, but chance forced-choice recognition for rsvp pictures. In *Proceedings of the Association for Research in Vision and Ophtalmology, ARVO*, Ft. Lauderdale, Florida.

VON DER MALSBURG, C. (1981). The correlation theory of brain function. Internal report, 81-2, Max-Planck-Institut für Biophysikalische Chemie, Postfach 2841, 3400 Göttingen, FRG. Reprinted in E. Domany, J.L. van Hemmen, and K. Schulten, editors, *Models of Neural Networks II*, chapter 2, pages 95–119. Springer, Berlin, 1994.

VON DER MALSBURG, C. AND BUHMANN, J. (1992). Sensory segmentation with coupled neural oscillators. *Biological Cybernetics*, 67:233–242.

VON DER MALSBURG, C. AND REISER, K. (1995). Pose invariant object recognition in a neural system. In *Proceedings of the International Conference on Artificial Neural Networks ICANN'95*, pages 127–132, Paris. EC2 & Cie.

VORBRÜGGEN, J. C. (1995). Data-driven segmentation of grey-level images with coupled nonlinear oscillators. In *Proceedings of the International Conference on Artificial Neural Networks ICANN'95*, pages 297–302, Paris. EC2 & Cie.

WISKOTT, L. (1995). *Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis*, volume 53 of *Reihe Physik.* Verlag Harri Deutsch, Thun, Frankfurt a. Main, Germany. PhD thesis.

WISKOTT, L., FELLOUS, J.-M., KRÜGER, N., AND VON DER MALSBURG, C. (1995). Face recognition and gender determination. In *Proceedings of the International Workshop on Automatic Face- and Gesture-Recognition, IWAFGR*, pages 92–97, Zurich.

WISKOTT, L. AND VON DER MALSBURG, C. (1993). A neural system for the recognition of partially occluded objects in cluttered scenes. *Int. J. of Pattern Recognition and Artificial Intelligence*, 7(4):935–948.

ZUCKER, R. S. (1989). Short-term synaptic plasticity. *Ann. Rev. Neuroscience*, 12:13–31.