In Sirosh, J., Miikkulainen, R., and Choe, Y., editors, Lateral Interactions in the Cortex: Structure and Function. The UTCS Neural Networks Research Group, Austin, TX. Electronic book, ISBN 0-9647060-0-8, http://www.cs.utexas.edu/users/nn/web-pubs/htmlbook96, Chapter 11 (1996)

Face Recognition by Dynamic Link Matching

Laurenz Wiskott and Christoph von der Malsburg*
Institut für Neuroinformatik
Ruhr Universität Bochum
D-44780 Bochum, Germany
{laurenz,malsburg}@neuroinformatik.ruhr-uni-bochum.de

Abstract

We present a neural system for the recognition of objects from realistic images, together with results of tests of face recognition from a large gallery. The system is inherently invariant with respect to shift, and is robust against many other variations, most notably rotation in depth and deformation. The system is based on Dynamic Link Matching. It consists of an image domain and a model domain, which we tentatively identify with primary visual cortex and infero-temporal cortex. Both domains have the form of neural sheets of hypercolumns, which are composed of simple feature detectors (modeled as Gabor-based wavelets). Each object is represented in memory by a separate model sheet, that is, a two-dimensional array of features. The match of the image to the models is performed by network self-organization, in which rapid reversible synaptic plasticity of the connections ("dynamic links") between the two domains is controlled by signal correlations, which are shaped by fixed inter-columnar connections and by the dynamic links themselves. The system requires very little genetic or learned structure, relying essentially on the rules of rapid synaptic plasticity and the *a priori* constraint of preservation of topography to find matches. This constraint is encoded within the neural sheets with the help of lateral connections, which are excitatory over short range and inhibitory over long range.

1 Introduction

The intracortical wiring pattern is a fascinating scientific subject, as it seems to hold the key to the function of the brain, or the part of it that we are accustomed to take most seriously. That wiring pattern is unnervingly close to being all-to-all. It has been speculated that signals from any cell in cortex can reach any other by crossing just three synapses. Although this seems to make sense for a system in which any two data items may have to contact each other, near-to-complete wiring seems to leave little room for all the specific structure that according to our present view of the brain resides in its connections. The experimental techniques of anatomy and neurophysiology are much too limited to give us more than gross principles of a cortical wiring pattern. These principles are to

*also Dept. of Computer Science and Section for Neurobiology, University of Southern California, Los Angeles, CA 90089 a very large extent summarized by speaking about receptive field structures, columnar organization, regular local interactions of the general type of difference-of-Gaussians and topographical connection patterns between areae. Beyond that we are in a dark continent, which may, for all we know, be dominated by randomness. More likely, however, it is structured by intricate learned patterns that are too variable from individual to individual and from place to place to ever become a possible subject of experimental enquiry. All we can hope to learn is the principles of organization by which they are formed.

We are presenting here a model for invariant object recognition, together with tests on human face recognition from a large gallery. The model may be relevant to the discussion at hand since it makes minimal assumptions about genetically generated connection patterns — certainly none that go beyond the principles enumerated — and relies largely on rapid reversible synaptic self-organization during the recognition process to create the much more specific connections required for a concrete recognition act. The model relies on Dynamic Link Matching (DLM) the qualitative principle of which has been described before (von der Malsburg, 1981; von der Malsburg, 1985; Konen and Vorbrüggen, 1993; Konen et al., 1993). The model described here goes beyond previously published versions in being more complete in its dynamic formulation, including mechanisms for autonomous activity blob dynamics, attention dynamics, and dynamic interaction between the stored models to implement the actual decision process during recognition.

A few words are in order to relate the jargon used in the description of our model to the biological background (the reader may want to come back to this "dictionary" while reading the next section). The term *image* refers to a cortical image domain which corresponds to the primary visual cortex V1 and possibly also to other areae up to perhaps V4. The image or image domain has the form of a graph. The nodes of the graph correspond to hypercolumns, that is, to collections of those feature specific neurons that are activated from one retinal point. In our system we formalize the activity of the sets of feature cells within hypercolumns as *jets*. As features we use Gabor-based wavelets. The links of the image graph correspond to lateral connections between nodes. An image on the retina selects a subset of the feature cells in the image domain. The selected neurons are then stochastically activated (these fluctuations not being driven by the visual signal). It is important that this stochastic activity takes a form that is characterized by temporal short-range correlations. These correlations express the neighborhood relations of visual features in the image and are produced by the lateral connections within the image domain. In our specific system the stochastic signal in the image domain (and also in the model domain) has the form of a local running blob of activity that is confined to an attention window. Apart from the local correlations the details of the activity process are not important, however.

The *models* (see right side of Figure 1) collectively form the model domain. We imagine this to be identified with some part of inferotemporal cortex. The *nodes of the models* again have the form of *jets* and are collections of neurons carrying feature labels. They are laterally connected much like nodes in the image domain. In our system the different models are totally disjoint. In the biological case models are likely to have partial overlap, in terms of single nodes or even partial networks. The stochastic activity process in the models is similar to that in the image domain, except for the interactions between models, which have the form of local co-operation (correlating activity between structurally corresponding points) and global competition between entire models.

The image domain and the model domain are bi-directionally connected by *dynamic links*. These correspond to connections between primary and infero-temporal cortex. These connections are assumed to be plastic on a fast time scale (changing radically during a single recognition event), this plasticity being reversible. The strength of a connection between any two nodes in the image and a model is controlled by the *jet similarity* between them, which roughly corresponds to the

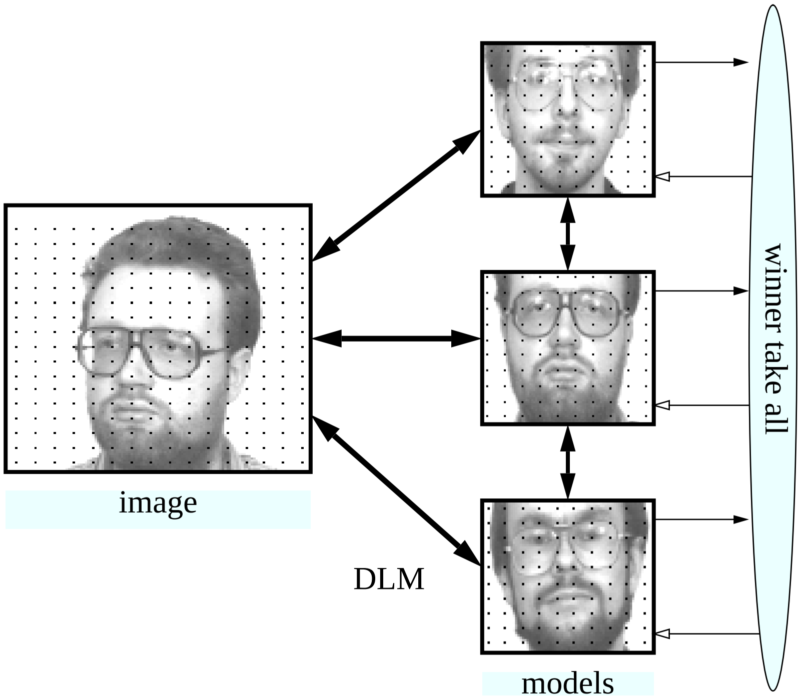

Figure 1: Architecture of the DLM face recognition system. Several models are stored as neural layers of local features on a 10×10 grid, as indicated by the black dots. A new image is represented by a 16×17 layer of nodes. Initially, the image is connected all-to-all with the models. The task of DLM is to find the correct mapping between the image and the models, providing translational invariance and robustness against distortion. Once the correct mapping is found, a simple winner-take-all mechanism can detect the model that is most active and most similar to the image.

number of features that are common to the two nodes.

2 The System

2.1 Architecture and Dynamics — Overview

Figure 1 shows the general architecture of the system. Faces are represented as rectangular graphs by layers of neurons. Each neuron represents a node and has a jet attached. A jet is a local description of the grey-value distribution based on the Gabor transform (see Lades et al., 1993; Wiskott et al., 1995). Topographical relationships between nodes are encoded by excitatory and inhibitory lateral connections. The model graphs are scaled horizontally and vertically and aligned manually, such that certain nodes of the graphs are placed on the eyes and the mouth (cf. Section 3.1). Model layers (10×10 neurons) are smaller than the image layer (16×17 neurons). Since the face in the image may be arbitrarily translated, the connectivity between model and image domain has to be all-to-all initially. The connectivity matrices are initialized using the similarities between the jets of the connected neurons. DLM serves as a process to restructure the connectivity matrices and to find the correct mapping between the models and the image (see Figure 2). The models cooperate with the image depending on their similarity. A simple winner-take-all mechanism sequentially rules out the least active and least similar models, and the best-fitting one eventually survives.

The dynamics on each layer is such that it produces a running blob of activity which moves continuously over the whole layer. An activity blob can easily be generated from noise by local

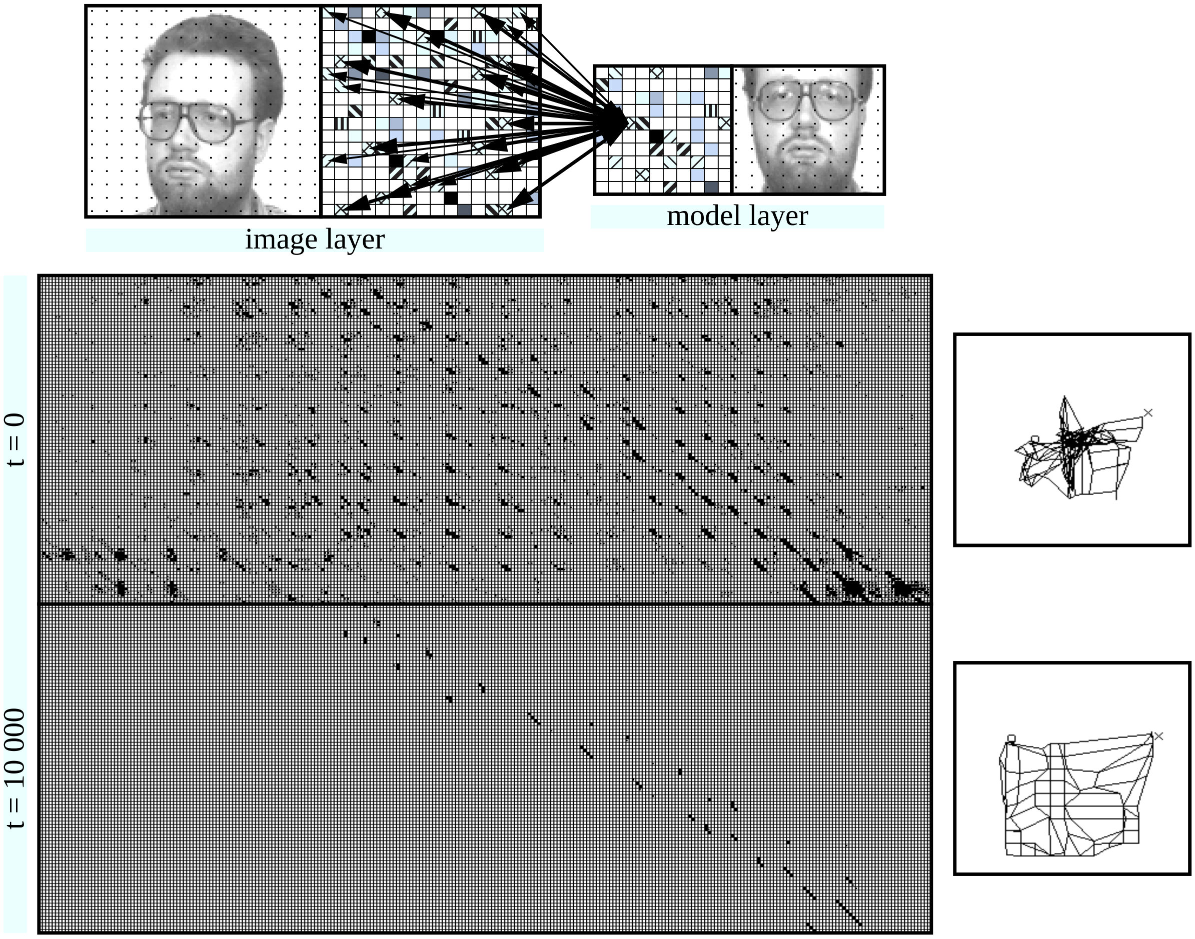

Figure 2: Initial and final connectivity for DLM. Image and model are represented by layers of 16×17 and 10×10 nodes respectively. Each node is labeled with a local feature indicated by small texture patterns. Initially, the image layer and the model layer are connected all-to-all with synaptic weights depending on the feature similarities of the connected nodes, indicated by arrows of different line widths. The task of DLM is to select the correct links and establish a regular one-to-one mapping. We see here the initial connectivity at $t = 0$ and the final one at $t = 10000$. Since the connectivity between a model and the image is a four-dimensional matrix, it is difficult to visualize it in an intuitive way. If the rows of each layer are concatenated to a vector, top row first, the connectivity matrix becomes two-dimensional. The model index increases from left to right, the image index from top to bottom. High similarity values are indicated by black squares. A second way to illustrate the connectivity is the net display shown at the right. 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 location of the nodes indicate the center 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. (see Figure 5 for connectivity development in time)

excitation and global inhibition. It is caused to move by delayed self-inhibition, which also serves as a memory for the locations where the blob has recently been. Since the models are aligned with each other, it is reasonable to enforce alignment between their running blobs by excitatory connections between neurons representing the same facial location. The blobs on the image and the model layers cooperate through the connection matrices; they tend to align and induce correlations between corresponding neurons. Then, fast synaptic plasticity and a normalization rule coherently modify the synaptic weights, and the correct connectivities between models and image layer can develop. Since the models get different input from the image, they differ in their total activity. The model with strongest connections from the image is the most active one. The models compete on the basis of their total activity. After a while the winner-take-all mechanism suppresses the least competitive models, and eventually only the best model survives. Since the image layer may be significantly larger than the model layers, we introduce an attention window in form of a large blob. It interacts with the running blob, restricts its region of motion, and can be shifted by it to the actual face position.

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; in order to make the description clearer, parts of the equations in Table 1 corresponding to these functions will be repeated.

2.2 Blob Formation

Blob formation on a layer of neurons can easily be achieved by local cooperation and global inhibition (Amari, 1977). Local cooperation generates clusters of activity, and global inhibition 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 cooperation and inhibition. The corresponding equations are (cf. Equations 1, 3, and 4):

$$\dot{h}_i(t) = -h_i + \sum_{i'} \left(g_{i-i'}\sigma(h_{i'})\right) - \beta_h \sum_{i'} \sigma(h_{i'}), \tag{8}$$

$$g_{i-i'} = \exp\left(-\frac{(i-i')^2}{2\sigma_g^2}\right), \tag{9}$$

$$\sigma(h) = \left\{ \begin{array}{ccl} 0 & : & h \leq 0 \\ \sqrt{h/\rho} & : & 0 < h < \rho \\ 1 & : & h \geq \rho \end{array} \right. . \tag{10}$$

The internal state of the neurons is denoted by h_i, where i is a two-dimensional Cartesian coordinate for the location of the neuron. The neurons are arranged on a regular square lattice with spacing 1, i.e., $i = (0,0), (0,1), (0,2), ..., (1,0), (1,1), ...$ The neural activity (which can be interpreted as a mean firing rate) is determined by the squashing function $\sigma(h)$ of the neuron's internal state h. The neurons are connected excitatorily through the Gaussian interaction kernel g. The strength of global inhibition is controlled by β_h. It is obvious that a blob can only arise if $\beta_h < g_0 = 1$ (imagine only one neuron is active), and that the blob is larger for smaller β_h. Infinite growth of h is prevented by the decay term $-h$, because it is linear, while the blob formation terms saturate due to the squashing function $\sigma(h)$. The special shape of $\sigma(h)$ is motivated by three factors. Firstly, σ vanishes for negative values to suppress oscillations in the simulations

Layer dynamics:

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

Attention dynamics:

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

Link dynamics:

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

Recognition dynamics:

$$
\begin{aligned}
r^p(t_0) &= 1 \\
\dot{r}^p(t) &= \lambda_r r^p \left(F^p - \max_{p'} (r^{p'} F^{p'}) \right) \qquad (7) \\
F^p(t) &= \sum_i \sigma(h_i^p)
\end{aligned}
$$

Table 1: Formulas of the DLM face recognition system

Variables:			
h			internal state of the layer neurons
s			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
$\mathcal{N}(\mathcal{J})$			saliency 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 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

by preventing undershooting. Secondly, the high slope for small arguments stabilizes small blobs and makes blob formation from low noise easier, because for small values of h the interaction terms dominate over the decay term. Thirdly, the finite slope region between low and high argument values allows the system to distinguish between the inner and outer parts of the blobs by making neurons in the center of a blob more active than at its periphery. Additional multiplicative parameters of the decay or cooperation terms would only change time and activity scale, respectively, and do not generate qualitatively new behavior. In this sense the parameter set is complete and minimal.

2.3 Blob Mobilization

Generating a running blob can be achieved by delayed self-inhibition, which drives the blob away from its current location; the blob generates new self-inhibition at the new location. This mechanism produces a continuously moving blob (see Figure 3). The driving force and the recollection time as to where the blob has been can be independently controlled by their respective time constants. The corresponding equations are (cf. Equations 1 and 2):

$$\begin{aligned}\dot{h}_i(t) &= -h_i + \sum_{i'} \left(g_{i-i'}\sigma(h_{i'})\right) - \beta_h \sum_{i'} \sigma(h_{i'}) - \kappa_{hs} s_i, &(11)\\ \dot{s}_i(t) &= \lambda_\pm (h_i - s_i). &(12)\end{aligned}$$

The self-inhibition s is realized by a leaky integrator with decay constant $\lambda_\pm$. The decay constant has two different values depending on whether $h - s$ is positive or negative. This accounts for the two different functions of the self-inhibition. The first function is to drive the blob forward. In this case $h > s$ and a high decay constant λ_+ is appropriate. The second function is to indicate where the blob has recently been, i.e., to serve as a memory and to repel the blob from regions recently visited. In this case $h < s$ and a low decay constant λ_- is appropriate. For small layers, λ_- should be larger than for large ones, because the blob visits each location more frequently. The speed of the blob is controlled by λ_+ and the coupling parameter κ_{hs}. They may also change the shape of the blob. Small values such as those used in our simulations allow the blob to keep its equilibrium shape and drive it slowly; large values produce a fast-moving blob distorted to a kidney-shape.

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 a connection matrix. 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 corresponding equation is (cf. Equation 1):

$$\begin{aligned}\dot{h}_i^p(t) &= -h_i^p + \sum_{i'} \left(g_{i-i'}\sigma(h_{i'}^p)\right) - \beta_h \sum_{i'} \sigma(h_{i'}^p) - \kappa_{hs} s_i^p \\ &\quad +\kappa_{hh} \max_j \left(W_{ij}^{pq}\sigma(h_j^q)\right), &(13)\\ \dot{s}_i^p(t) &= \lambda_\pm (h_i^p - s_i^p). &(14)\end{aligned}$$

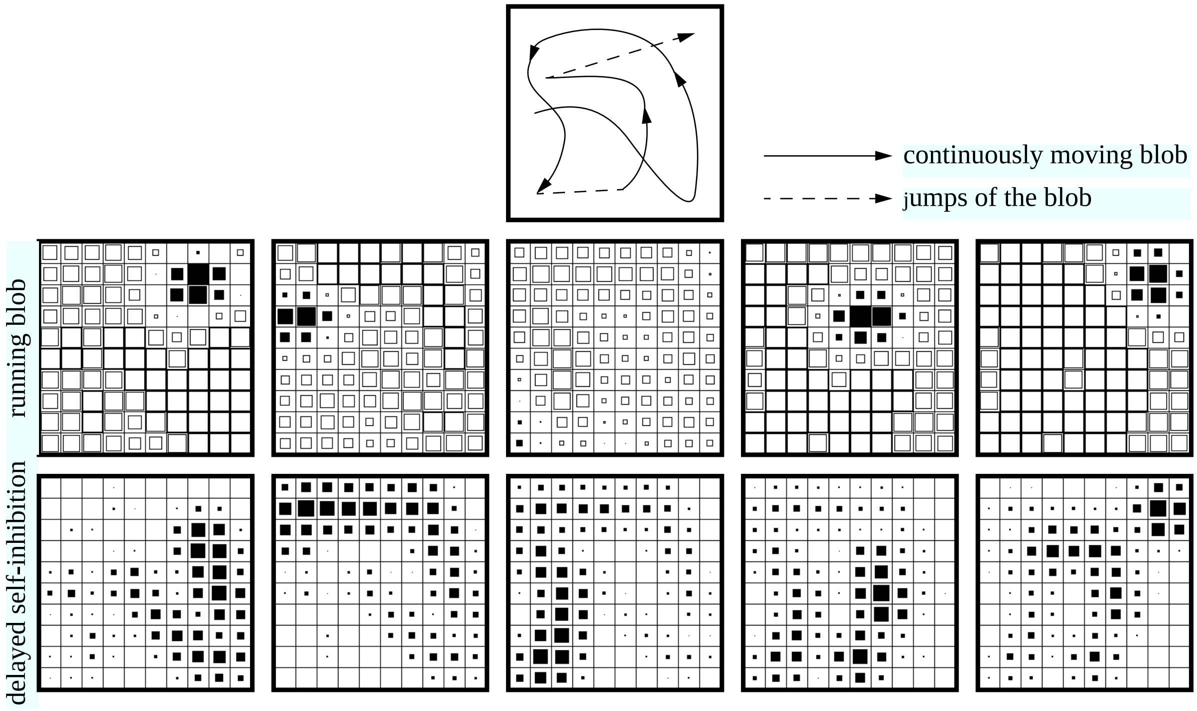

Figure 3: A sequence of layer states as simulated with Equations 11 and 12. The activity blob h shown in the middle row has a size of approximately six active nodes and moves continuously over the whole layer. Its course is shown in the upper diagram. The delayed self-inhibition s, shown in the bottom 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.)

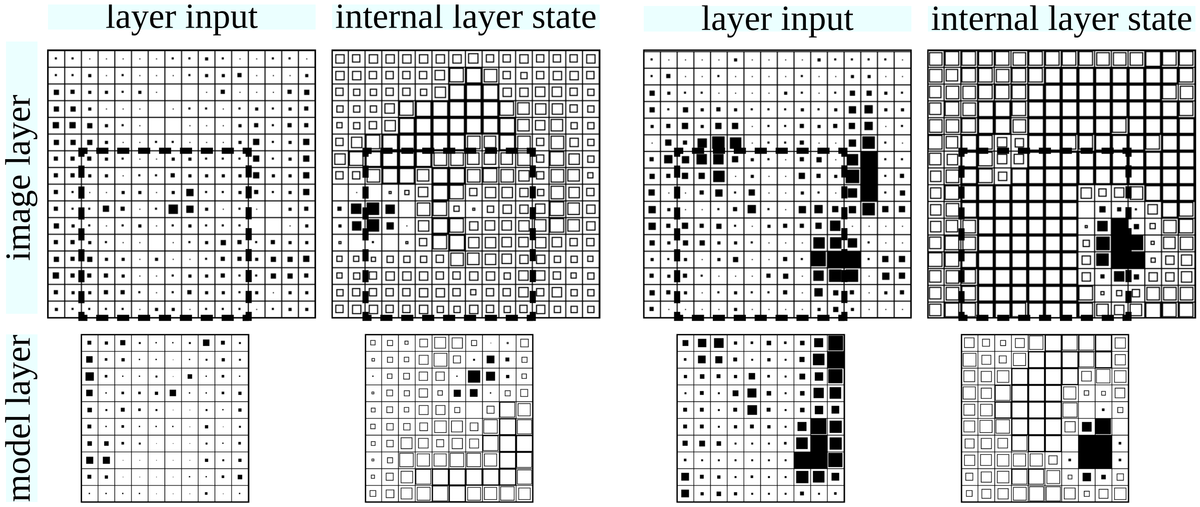

Figure 4: Synchronization between two running blobs as simulated with Equations 13 and 14. Layer input as well as the internal layer state h is shown at an early stage, in which the blobs of two layers are not yet aligned, left, and at a later state, right, when they are aligned. The two layers are of different size, and the region in layer 1 that 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.

The two layers are indicated by the indices p and q. The synaptic weights of the connections are W, and the strength of mutual interaction is controlled by the parameter κ_{hh}. (The reason why we use the maximum function instead of the usual sum will be discussed in Section 2.10.)

2.5 Link Dynamics

One principle of DLM is that the links between two layers can be cleaned up and structured on the basis of correlations between pairs of neurons (see Figure 5 and Movie 1). The correlations result from the layer synchronization described in the previous section. The link dynamics typically consists of a growing 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 which suppresses all others. The corresponding equations are (cf. Equations 6):

$$\begin{aligned} W_{ij}^{pq}(t_0) &= \mathcal{S}_{ij}^{pq} = \max\left(\mathcal{S}_\phi(\mathcal{J}_i^p, \mathcal{J}_j^q), \alpha_\mathcal{S}\right), \\ \dot{W}_{ij}^{pq}(t) &= \lambda_W \left(\sigma(h_i^p)\sigma(h_j^q) - \Theta\left(\max_{j'}(W_{ij'}^{pq}/\mathcal{S}_{ij'}^{pq}) - 1\right)\right) W_{ij}^{pq}. \end{aligned} \qquad (15)$$

Links are initialized by the similarity $\mathcal{S}_\phi$ between the jets $\mathcal{J}$ of connected nodes (see Wiskott, 1995). The parameter $\alpha_\mathcal{S}$ guarantees a minimal positive synaptic weight, permitting each link to suppress others, even if the similarity between the connected neurons is small. This can be useful to obtain a continuous mapping if a link has a neighborhood of strong links inducing high correlations between the pre- and postsynaptic neurons of the weak link. The synaptic weights grow exponentially, controlled by the correlation between connected neurons defined as the product

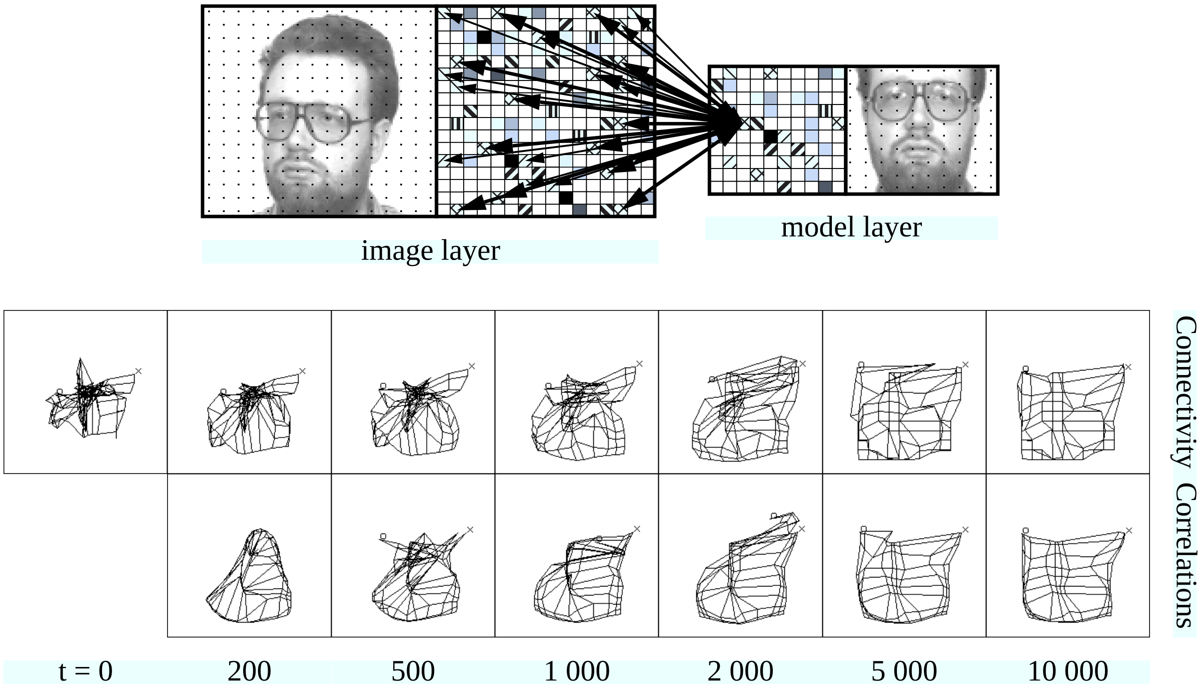

Figure 5: Connectivity and correlations developing in time. It can be seen how the correlations develop faster and are cleaner than the connectivity. Both are iteratively refined on the basis of the other.

of their activities $\sigma(h_i^p)\sigma(h_j^q)$. The learning rate is additionally controlled by λ_W. Due to the Heavyside-function Θ, normalization takes place only if links grow beyond their initial value. Then, the link dynamics is dominated by the normalization term, with a common negative contribution for all links converging to the same neuron. Notice that the growth term, based on the correlation, is different for different links. Thus the link with the highest average correlation will eventually suppress all others converging to the same neuron. Since the similarities $\mathcal{S}_\phi$ cannot be larger than 1, the synaptic weights W are restricted to the interval $[0, ..., 1]$.

Movie 1: Connectivity and correlations developing in time as in Figure 5 (//http:...., 350 kB).

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 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. 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, but it can be shifted by the latter into a region where input is especially large and favors activity. The attention blob therefore automatically aligns with the actual face position (see Figures 6, 7 and Movie 2). The attention blob layer is initialized with a primitive segmentation cue, in this case the norm of the respective jets (see Wiskott, 1995), since the norm indicates the presence of textures of high contrast. The corresponding equations are (cf. Equations 1 and 5):

$$\begin{aligned}\dot{h}_i^p(t) &= -h_i^p + \sum_{i'} \left(g_{i-i'}\sigma(h_{i'}^p)\right) - \beta_h \sum_{i'} \sigma(h_{i'}^p) - \kappa_{hs} s_i^p \\ &\quad + \kappa_{hh} \max_j \left(W_{ij}^{pq}\sigma(h_j^q)\right) + \kappa_{ha}\left(\sigma(a_i^p) - \beta_{ac}\right), \qquad (16)\\ \dot{s}_i^p(t) &= \lambda_\pm(h_i^p - s_i^p), \qquad (17)\end{aligned}$$

$$\begin{aligned}a_i^p(t_0) &= \alpha_{\mathcal{N}}\mathcal{N}(\mathcal{J}_i^p), \\ \dot{a}_i^p(t) &= \lambda_a\left(-a_i^p + \sum_{i'} g_{i-i'}\sigma(a_{i'}^p) - \beta_a \sum_{i'} \sigma(a_{i'}^p) + \kappa_{ah}\sigma(h_i^p)\right). \qquad (18)\end{aligned}$$

The equations show that the attention blob a is generated by the same dynamics as was discussed in Section 2.2 for the formation of the running blob without delayed self-inhibition, though since the attention blob is to be larger than the running blob, β_a has to be smaller than β_h. The attention blob restricts the region for the running blob via the term $\kappa_{ha}\left(\sigma(a_i^p) - \beta_{ac}\right)$, which is an 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. By this means the running blob can slowly shift the attention blob into its favored region. The dynamics of the attention blob has to be slower than that of the running blob; this is controlled by a value $\lambda_a < 1$. $\mathcal{N}$ is the norm of the jets, and $\alpha_{\mathcal{N}}$ determines the initialization strength.

2.7 Recognition Dynamics

Each model cooperates with the image depending on its similarity. The most similar model cooperates most successfully and is the most active one. Hence, the total activity of the model layers indicates which is the correct one. 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. The corresponding equations are (cf. Equations 1 and 7):

$$\begin{aligned}\dot{h}_i^p(t) &= -h_i^p + \sum_{i'} \left(g_{i-i'}\sigma(h_{i'}^p)\right) - \beta_h \sum_{i'} \sigma(h_{i'}^p) - \kappa_{hs} s_i^p \qquad (19)\\ &\quad + \kappa_{hh} \max_j \left(W_{ij}^{pq}\sigma(h_j^q)\right) + \kappa_{ha}\left(\sigma(a_i^p) - \beta_{ac}\right) - \beta_\theta \Theta(r_\theta - r^p), \\ \dot{s}_i^p(t) &= \lambda_\pm(h_i^p - s_i^p), \qquad (20)\end{aligned}$$

$$\begin{aligned}r^p(t_0) &= 1, \\ \dot{r}^p(t) &= \lambda_r r^p \left(F^p - \max_{p'}(r^{p'}F^{p'})\right), \qquad (21)\\ F^p(t) &= \sum_i \sigma(h_i^p).\end{aligned}$$

The total layer activity is considered as a fitness F^p, different for each model p. The modified evolution equation can be easily analyzed if the F^p are assumed to be constant in time and the recognition variables r^p are initialized to 1. For the model layer p_b with the highest fitness, the equation simplifies to $\dot{r}^{p_b}(t) = \lambda_r r^{p_b}(1 - r^{p_b})F^{p_b}$ with a stable fixed point at $r^{p_b} = 1$. For all other models the equation then simplifies to $\dot{r}^p(t) = \lambda_r r^p(F^p - F^{p_b})$, which results in an exponential decay

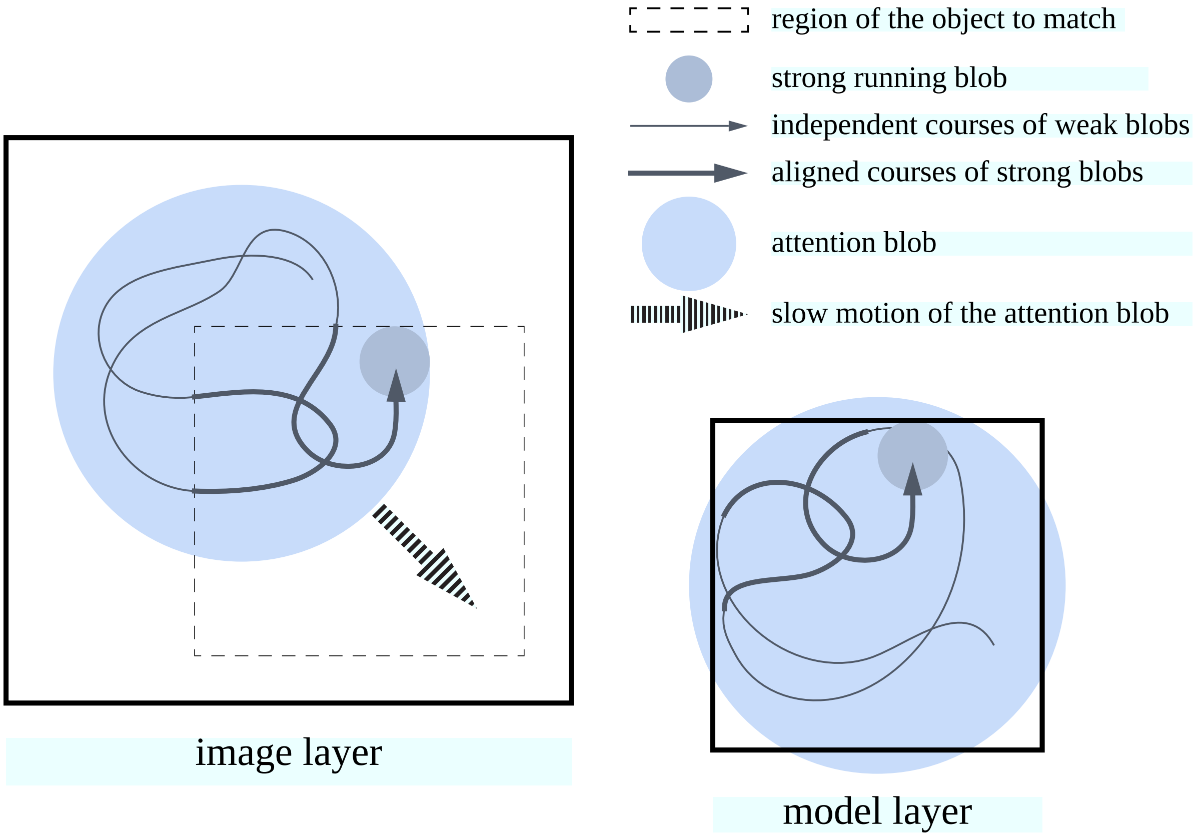
region of the object to match
strong running blob
independent courses of weak blobs
aligned courses of strong blobs
attention blob
slow motion of the attention blob
image layer
model layer

Figure 6: Schematic of the attention blob's function. The attention blob restricts the region in which the running blob can move. The attention blob, on the other hand, receives input from the running blob. That input will be strong in regions where the blobs in both layers cooperate and weak where they do not (see Figure 4). Due to this interaction the attention blob slowly moves to the correct region indicated by the square made of dashed lines. The attention blob in the model layer is required to keep the conditions for the running blobs symmetrical.

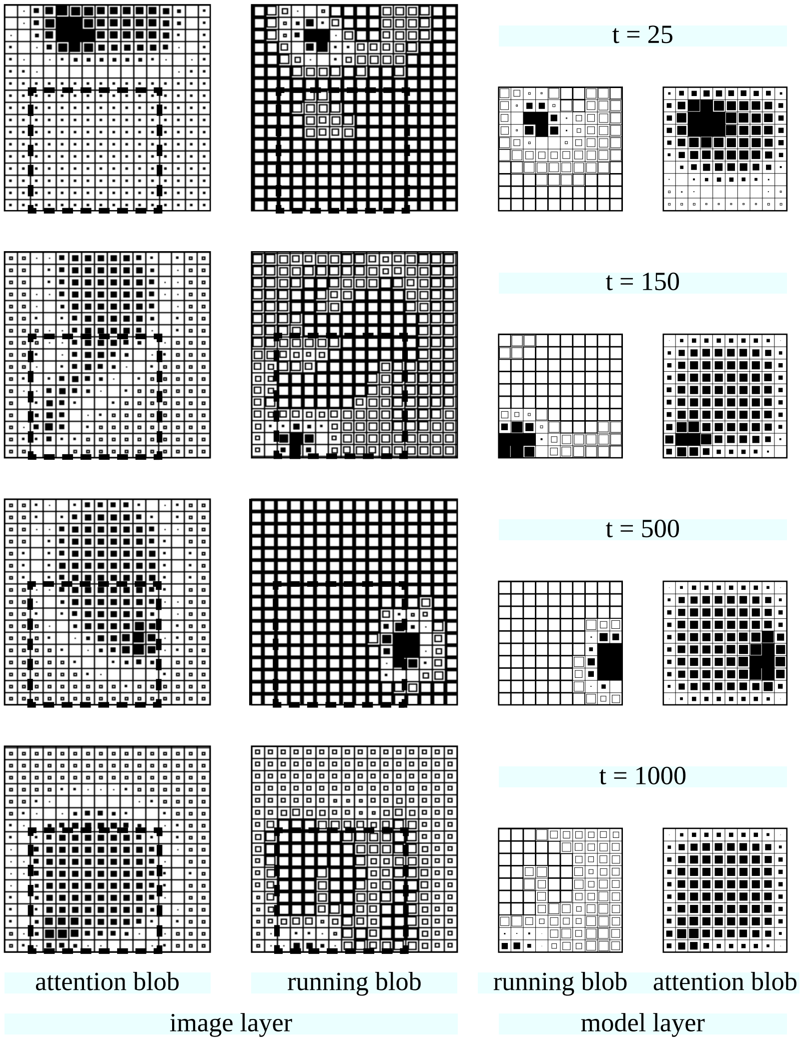

Figure 7: Function of the attention blob, using an extreme example of an initial attention blob manually misplaced for demonstration. At $t = 150$ the two running blobs ran synchronously for a while, and the attention blob has a long tail. The blobs then lost alignment again. From $t = 500$ on, the running 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 (cf. Movie 2). Without an attention blob the two running blobs may synchronize sooner, but the alignment will never become stable (see Movie 3).

Movie 2: Attention blob dynamics as in Figure 7. Shown are the running blob (black), the delayed self-inhibition (red), and the attention blob (blue) on image and model layer (//http:...., 336 kB).

Movie 3: Blob dynamics as in Movie 2, but without attention blobs, demonstrating that alignment does not get stable (//http:...., 196 kB).

of the r^p for all $p \neq p_b$. When a recognition variable r^p drops below the suppression threshold r_θ, the activity on layer p is suppressed by the term $-\beta_\theta \Theta(r_\theta - r^p)$. The time scale of the recognition dynamics can be controlled by λ_r.

2.8 Bidirectional Connections

The connectivity between two layers is bidirectional and not unidirectional as in the previous system (Konen and 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 and leads to the term $+\sum_{i'} \max_{p'} \left(g_{i-i'} \sigma(h_{i'}^{p'})\right)$ instead of $+\sum_{i'} (g_{i-i'} \sigma(h_{i'}^{p}))$ 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 very 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 very 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 can 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 very different and without a common structure, as it is for general objects. The disadvantage is that the system needs much more time to decide which model to accept, because the relative layer activities in the beginning depend much more on chance than in the other setup.

2.10 Maximum Versus Sum Neurons

The model neurons used here use the maximum over all input signals instead of the sum. The reason is that the sum would mix up many different signals, while only one can be the correct one, i.e., the total input would be the result of one correct signal and many misleading ones. Hence the signal-to-noise ratio would be very low. We have observed an example where even a model identical to the image was not picked up 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. Of 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 a 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 1). Though the grids are rectangular and regular, i.e., the spacing between the nodes is constant for each dimension, the graphs are scaled horizontally in the x- and vertically in the y-direction 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-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 W^a was derived by taking the maximum synaptic weight over all real models for each link:

$$\begin{aligned} W^a_{ij}(t_0) &= \max_{pq} W^{pq}_{ij}(t_0), \\ \dot{W}^a_{ij}(t) &= 0. \end{aligned} \tag{22}$$

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 translational invariance. Full rotational invariance is lost, but the jets used are not rotationally invariant in any case. 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 that 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 strong tendency to stay in the center.

3.3 Results

Figure 8 shows two recognition examples, one using a test face rotated in depth and the other using a face with very different expression. In both cases the gallery contains five models. Due to the tight connections between the models, the layer activities show the same variations and differ only very 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 examples were monitored for 2000 units of simulation time. An attention phase of 1000 time units had been applied before, but is not shown here. The second recognition task was obviously harder than the first. 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 many 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 (the examples in Figure 8 are not typical in this respect). It is interesting to consider recognition times. 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 for visual object recognition we are presenting here marks the extreme end of a scale, relying minimally on pre-existing structure. In fact, all it needs is some natural intracortical connection patterns, one stored example for each object to be recognized, and a simple mechanism of on-line self-organization in the form of rapid reversible synaptic plasticity. This distinguishes it from many alternative neural models for object recognition, which require extensive control structures (Anderson and Van Essen, 1987) or specific feature hierarchies, to be created by training (Fukushima et al., 1983; LeCun et al., 1989), before the first object can be recognized. The lateral connections within the image domain and the model domain of our system encode the *a priori* constraint of conservation of spatial continuity during the match. The match itself is realized with the help of the rapid self-organization of the synaptic connections between image and models. This self-organization is controlled by signal correlations and by feature similarity between image points and model points. For each object to be recognized just a single model needs to be stored, which can be done with the help of simple mechanisms of associative memory (von der Malsburg, 1988). (For the accommodation of substantial rotation in depth the object needs to

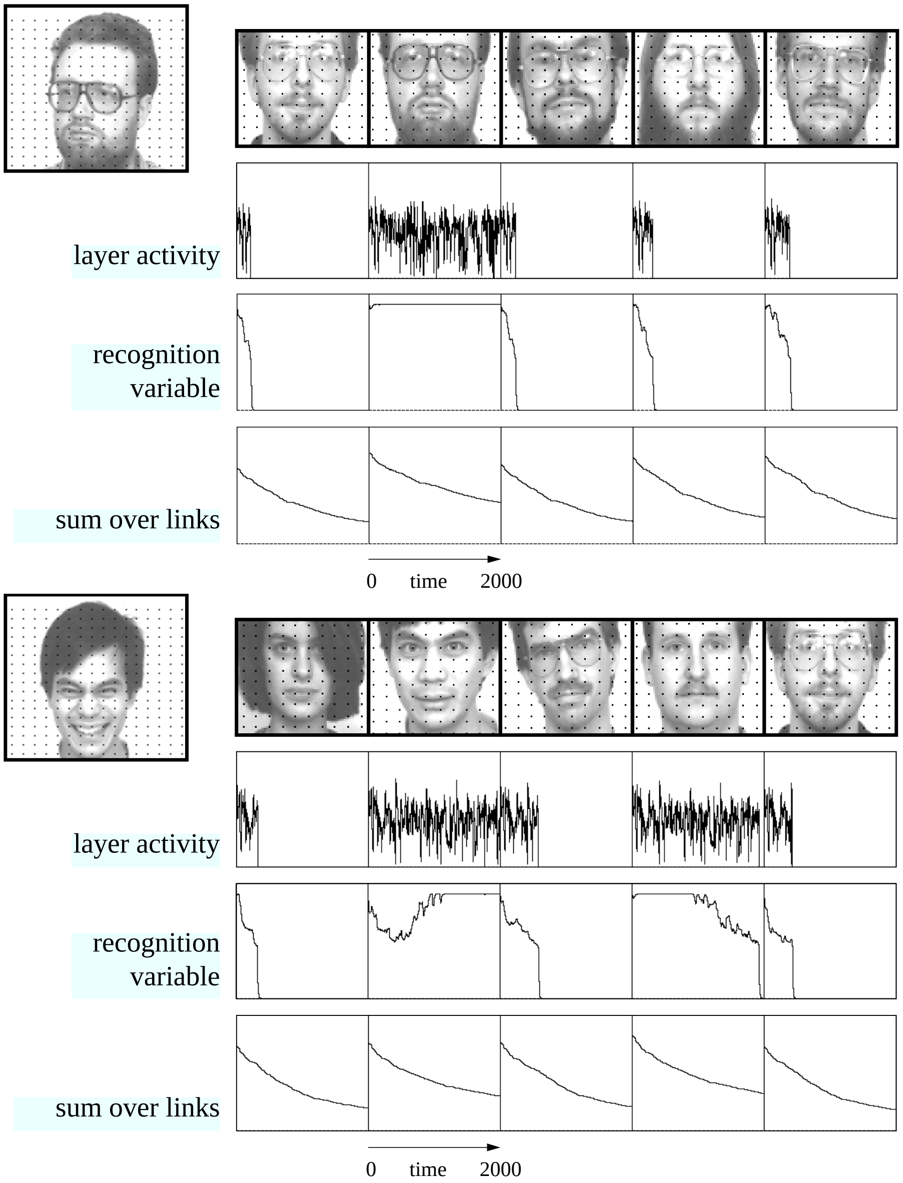

Figure 8: Simulation examples of DLM recognition. The test images are shown on the left with 16×17 neurons indicated by black dots. The models have 10×10 neurons and are aligned with each other. The respective total layer activities, i.e., the sum over all neurons of one model, are shown in the upper graphs. 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.

be inspected from many angles and the resulting models need to be fused into one model graph, a principle demonstrated by Reiser (1991).) From these properties of our system results a very clear-cut message concerning the issue of intracortical connections: visual object recognition can be understood on the basis of simple connectivity structures and mechanisms of plasticity that are already known today or at least are well within the reach of existing experimental techniques!

Our model leaves open a number of questions regarding the structure of lateral connections, especially in the model domain. The global interaction between models (our "recognition dynamics") could be realized with the help of a single cardinal cell per model, or it could take the form of a distributed set of connections between model neurons (as formulated in von der Malsburg, 1988). More work is required to decide this issue. The anatomy of the local interaction between models, second term on the right-hand side of Equation 1, can only be discussed after the relative anatomical placement of different models has become clear. Also, the extent and the nature of the overlap between models in terms of common neurons and common connections must be clarified first. Two extreme versions are imaginable, i) models are laid down in mutual register in terms of internal position, and ii) there is a fixed spatial array of feature types in infero-temporal cortex (for which there is faint experimental evidence (Tanaka, 1993)), and laying down a model consists in selecting appropriate feature cells and connecting these as required by the inner structure of the model. In the first case, the lateral model connections would be tidy and local within the cortical tissue (at least their excitatory part), in the second they would form a diffuse fiber plexus without any apparent anatomical structure. A further aspect of intracortical connectivity that we are totally ignoring in the present system concerns intra-hypercolumnar connectivity. This is implicitly present, being required to organize the necessary feature specificity, and probably also for the evaluation of the feature similarity between a pair of hypercolumns ("nodes") in image and model.

Last, and by no means least, we have given short shrift to the issue of inter-areal organization of connections, by lumping all primary areae into one image domain and all infero-temporal areae into one model domain. Within the image domain, two extreme views could be taken. i) The different areae (V1, V2, V4, for instance) represent different mixtures of feature specificities and are tied together by rigid retinotopically organized connections. In that case areal structure could be ignored for the purposes of our present system, and neurons in different areae but subserving the same retinal point could just be lumped together into one "hyper-hypercolumn." ii) The synaptic projection systems between areae are substantially reorganized during the recognition process, areae perhaps forming sequential layers connecting V1 indirectly with IT, as proposed in (Anderson and Van Essen, 1987). Perhaps such an indirect connectivity scheme can reduce the enormous number of fibers required by our system for connecting any pair of points in image and models.

There is one apparent mismatch between our system and the reality of object recognition in the brain of adults: the time taken by the process. There are reports that objects of different type can be distinguished by human subjects in less than a tenth of a second (Subramaniam et al., 1995). In contrast, our system requires for the process many hundred sequential steps. It is not easy to interpret these sequential steps in terms of biological real time. The essential parameter seems, however, to be the temporal resolution with which signal correlations can be evaluated in our brain. This issue is at present under heated discussion (Softky, 1995; Shadlen and Newsome, 1995), but there is little hope that this resolution is better than one or a few milliseconds. In this case the hundreds of sequential steps required by our system translate into many hundred milliseconds, which is unrealistically long. Dynamic Link Matching needs this time to reduce the enormous ambiguity in the feature similarities between image and object points to a sparse set of connections between corresponding points. If this ambiguity could be decisively reduced with the help of highly specific feature types (which in an extreme case were private to one object type), recognition time could be cut drastically. The feature types we are using, Gabor-based wavelets, are very general and unspecific. It is likely that highly specific features can only be generated by a learning mechanism. It is our view that the basic mechanism of our system is used by the young animal to store and recognize objects early in its life. At first, each recognition process may take seconds, but the mechanism can be the basis for very efficient learning of specific feature types, a process that due to the Dynamic Link Mechanism is not hampered by confusion between different objects.

The most encouraging aspect of our system is its evident capability to solve the invariant object recognition problem in spite of all the difficulties and adversities posed by real images and in spite of large numbers and great structural overlap of objects to be distinguished. This puts it in sharp contrast to proposed recognition mechanisms that work only on simple toy examples. We therefore feel that this system is a foot in the door, and its remaining difficulties can be solved gradually. What is important in the context of the present book is the light our system sheds on the functional role of lateral connections in visual cortex.

Acknowledgements

This work has been funded by grants from the German Federal Ministry of Science and Technology (413-5839-01 IN 101 B/9) and from AFOSR (F49620-93-1- 0109). The computer simulations have been made using NSL (Neural Simulation Language) developed by Alfredo Weitzenfeld.

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.

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.

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

Fukushima, K., Miyake, S., and Ito, T. (1983). Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Transactions 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.

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., Maurer, T., and von der Malsburg, C. (1994). A fast dynamic link matching algorithm for invariant pattern recognition[1]. *Neural Networks*, 7(6/7):1019–1030.

Konen, W. and Vorbrüggen, J. C. (1993). Applying dynamic link matching to object recognition in real world images[2]. In Gielen, S. and Kappen, B., editors, *Proceedings of the International Conference on Artificial Neural Networks, ICANN*, pages 982–985, London. Springer-Verlag.

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

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.

Reiser, K. (1991). Learning persistent structure. Doctoral thesis, res. report 584, Hughes Aircraft Co., 3011 Malibu Canyon Rd. Malibu, CA 90265.

Shadlen, M. N. and Newsome, W. T. (1995). Is there a signal in the noise? *Current Opinion in Neurobiology*, 5:248–250.

Softky, W. (1995). Simple codes vs. efficient codes. *Current Opinion in Neurobiology*, 5:239–247.

Subramaniam, S., Biederman, I., Kalocsai, P., and Madigan, S. (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.

Tanaka, K. (1993). Neuronal mechanisms of object recogntition. *Science*, 262:685–688.

[1] http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/PUBLICATIONS/ABSTRACTS/KonMauMal94.html

[2] http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/PUBLICATIONS/ABSTRACTS/KonVor93.html

[3] http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/PUBLICATIONS/ABSTRACTS/LadVorBuh93.html

von der Malsburg, C. (1981). The correlation theory of brain function[4]. Internal report, 81-2, Max-Planck-Institut für Biophysikalische Chemie, Postfach 2841, 3400 Göttingen, FRG.

von der Malsburg, C. (1985). Nervous structures with dynamical links[5]. *Ber. Bunsenges. Phys. Chem.*, 89:703–710.

von der Malsburg, C. (1988). Pattern recognition by labeled graph matching[6]. *Neural Networks*, 1:141–148.

Wiskott, L. (1995). *Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis*[7]. PhD thesis, Fakultät für Physik und Astronomie, Ruhr-Universität Bochum, D-44780 Bochum.

Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C. (1995). Face recognition and gender determination[8]. In *Proceedings of the International Workshop on Automatic Face- and Gesture-Recognition, IWAFGR 95*, pages 92–97, Zurich.

[4] http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/PUBLICATIONS/ABSTRACTS/Mal81.html

[5] http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/PUBLICATIONS/ABSTRACTS/Mal85b.html

[6] http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/PUBLICATIONS/ABSTRACTS/Mal88c.html

[7] http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/PUBLICATIONS/ABSTRACTS/Wis95a.html

[8] http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/PUBLICATIONS/ABSTRACTS/WisFelKrue95.html