PERGAMON

Neural Networks 14 (2001) 345–354

Neural Networks

www.elsevier.com/locate/neunet

Contributed article

Graph matching vs mutual information maximization for object detection ☆

Ladan B. Shams[a,*], Mark J. Brady[b], Stefan Schaal[c,d]

[a]California Institute of Technology, Computation and Neural Systems, Division for Biology, MC 139-74, Pasadena, CA 92215, USA
[b]3M Corporate Research Laboratories, 3M Center, Building 235-3F-08, St. Paul, MN 55144, USA
[c]University of Southern California, Computer Science & Neuroscience, HNB 103, Los Angeles, CA 90089-2520, USA
[d]ERATO Kawato Dynamic Brain Project (JST), 2-2 Hikaridai, Seika-cho, Soraku-gun, 619-02 Kyoto, Japan

Received 28 December 1999; accepted 30 October 2000

Abstract

Labeled Graph Matching (LGM) has been shown successful in numerous object vision tasks. This method is the basis for arguably the best face recognition system in the world. We present an algorithm for visual pattern recognition that is an extension of LGM ('LGM$^+$'). We compare the performance of LGM and LGM$^+$ algorithms with a state of the art statistical method based on Mutual Information Maximization (MIM). We present an adaptation of the MIM method for multi-dimensional Gabor wavelet features. The three pattern recognition methods were evaluated on an object detection task, using a set of stimuli on which none of the methods had been tested previously. The results indicate that while the performance of the MIM method operating upon Gabor wavelets is superior to the same method operating on pixels and to LGM, it is surpassed by LGM$^+$. LGM$^+$ offers a significant improvement in performance over LGM without losing LGM's virtues of simplicity, biological plausibility, and a computational cost that is 2–3 orders of magnitude lower than that of the MIM algorithm.

Keywords: Pattern recognition; Object recognition; Graph matching; Mutual information maximization; Object detection; Lateral excitation; Image entropy; Gabor wavelets

1. Introduction

Computer vision of objects and faces comprises a multitude of tasks: detection, recognition, alignment, pose estimation, scene analysis, tracking, etc. Most algorithms developed in this realm are designed for one of these tasks and rarely generalize to others. For instance, the neural network developed by Rowley, Baluja, & Kanade, 1995 is highly successful in detecting faces in scenes, but not capable of recognition, and the face recognition algorithms developed in Pentland's laboratory (Turk & Pentland, 1991) are successful for face recognition but are not as apt at locating them.

However, the pattern recognition tasks mentioned above are highly related and, from a biological perspective, it seems likely that the brain uses mechanisms in these modules that are based on the same principles and rely on the same basic type of object representation. An algorithm that has been applied to all of the aforementioned vision tasks is Labeled Graph Matching (LGM) (von der Malsburg, 1988). LGM represents each pattern as a labeled graph where each node is labeled with a feature and the links between nodes encode topological relationships. Patterns are extracted and recognized by means of finding the optimal correspondence between two graphs. The implementations of LGM have primarily used responses of a family of Gabor wavelets as the label for each node, representing the local gray-level distribution of the image. This approach has had considerable success in various domains, including object detection and recognition (Konen & Vorbrüggen, 1993; Konen, Maurer, & von der Malsburg, 1994), face detection and recognition (Lades, Vorbrüggen, Buhmann, Lange, Malsburg, Würtz, & Konen et al., 1993; Wiskott, Fellous, Krüger, & von der Malsburg, 1997; Wiskott & von der Malsburg, 1995), gender determination (Wiskott, Fellous, Krüger, & von der Malsburg, 1995), scene analysis (Wiskott, 1996a,b; Wiskott & von der Malsburg, 1993), pose estimation (Krüger, Pötzsch, & von der Malsburg, 1997), face rotation (Maurer & von der Malsburg, 1995), tracking (Maurer & von der Malsburg, 1996), and object shape feature learning (Shams & von der Malsburg, 1999).

☆ This work was supported by NSF's IMSC program at USC, by NSF award no. 9710312, the ERATO Kawato Dynamic Brain Project, and the ATR Human Information Processing Labs.

* Corresponding author. Tel.: +1-626-395-2362; fax: +1-626-844-4514.
E-mail address: ladan@caltech.edu (L.B. Shams).

0893-6080/01/$ - see front matter
PII: S0893-6080(00)00099-X

Recently, there has been a surge of statistical approaches to the problems of pattern recognition, mostly based on information theoretic notions, such as entropy (Viola, 1995), Shannon information (Becker, 1995), and description length (Bienenstock, Geman, & Potter, 1997). A statistical approach which has enjoyed much attention in the past few years and has been adopted by several different groups is Mutual Information Maximization (some examples are (Collignon, Maes, Delaere, Vandermeulen, Suetens, & Marchal, 1995; McGarry, Jackson, Plantec, Kassell, & Downs, 1997; Moskalik, Carson, Meyer, Fowlkes, Rubin, & Rubidoux, 1995; Pluim, Maintz, & Viergever, 2000; Studholme, Hill, & Hawkes, 1996; Viola, 1995; Viola & Wells, 1995)).

In contrast to the solid statistical foundation of these methods, LGM has neither been derived from statistical principals nor does it explicitly exploit statistical image properties—LGM was designed based on biological plausibility, and invariant recognition constraints. An open question, thus, becomes how well LGM performs in comparison to these state-of-the-art statistical methods.

In this paper, we will introduce a new version of LGM that exploits the topology of an image through lateral interactions. For comparison, we adapt Viola's mutual information maximization (MIM) approach (Viola, 1995; Viola & Wells, 1995) to operate on the same Gabor wavelets as LGM. Viola's method was chosen for comparison as it is the pioneering work for a number of research efforts undertaken by various groups using the same method, and has been applied successfully to real images and to several pattern recognition applications. To render the study as objective as possible, we chose a task that both methods have primarily been applied to: object detection. In the remaining of the paper, we first describe the image representations we employ. Second, we introduce our new LGM method, or LGM$^+$, and the adapted version of Viola's Mutual Information approach. We will proceed by describing the pattern recognition task used for the study, and conclude with an empirical evaluation of the three methods: MIM, LGM, and LGM$^+$.

2. Image representation

The image representation we use is modeled after the feature detectors found in the mammalian primary visual cortex (V1). Each image is coded by a fixed grid of feature vectors modeled after cortical hypercolumns (Fig. 1). Given an image with gray values $W_{\boldsymbol{\mu}}$[1] defined on a two-dimensional lattice $\Im$ of pixel positions $\boldsymbol{\mu} \in \Im$, the Gabor

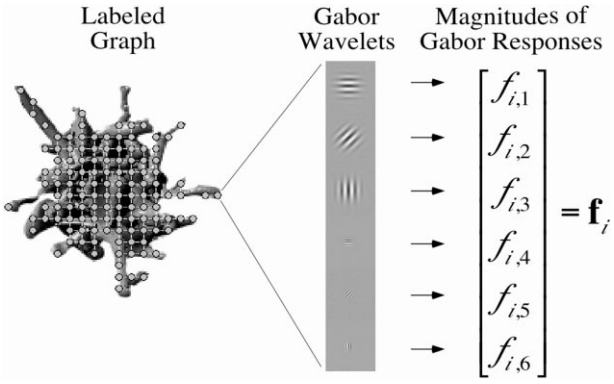

Fig. 1. Sketch of an image with a superimposed labeled graph. Each graph node is 'labeled' with a Gabor Jet, a vector of magnitudes of Gabor wavelet filter responses of different frequencies and orientations. The Gabor Jet can be conceived of as a computational model of cortical hypercolumns. Gabor wavelets of three different orientations at two different spatial frequencies are displayed in the right side of the figure. In our implementation, we use four orientations at three spatial frequencies.

transform, oriented along the vector $\mathbf{k}$, is given as

$$J_{\mathbf{k}\boldsymbol{\mu}} = \sum_{\boldsymbol{\mu}'} W_{\boldsymbol{\mu}'} \psi_{\mathbf{k}}(\boldsymbol{\mu} - \boldsymbol{\mu}'), \qquad \boldsymbol{\mu}, \boldsymbol{\mu}' \in \Im \qquad \text{with}$$
$$\psi_{\mathbf{k}}(\boldsymbol{\mu}) = \frac{\mathbf{k}^T\mathbf{k}}{\sigma^2} e^{-\frac{\mathbf{k}^T\mathbf{k}\,\boldsymbol{\mu}^T\boldsymbol{\mu}}{2\sigma^2}} \left(e^{i\mathbf{k}^T\boldsymbol{\mu}} - e^{-\sigma^2/2}\right) \tag{1}$$

where the kernel $\psi_{\mathbf{k}}(\boldsymbol{\mu})$ is a Gabor wavelet (Grossmann & Morlet, 1985). For each single point of the visual field, we assume a sampling of the frequency domain at three frequency levels, F, and, within a frequency level, at a set of four orientations, O, resulting in a 12-dimensional feature vector $\boldsymbol{f} \in F \times O$ (Fig. 1), called a 'jet' (Buhmann, Lange & von der Malsburg, 1989). Jets are conceived of as a simple model of hypercolumnar activity of visual cortex. As in biological complex cells, only magnitude[2] of Gabor responses are used (von der Malsburg, Shams, & Eysel, 1998).

3. Algorithms for matching

3.1. Labeled graph matching

A visual pattern can be represented via a graph containing nodes labeled with local features and links encoding the topological relationship between the features (Bienenstock & von der Malsburg, 1987). Based on this representation, the problem of pattern recognition can be formulated as labeled graph matching (LGM), where the goal is to find the one-to-one correspondence between the nodes of an input graph and those of a stored graph (Bienenstock & von der Malsburg, 1987). A good correspondence is one that respects the topological relationships between the

[1] In our notation, matrices are denoted by bold-face capitalized letters, e.g., **M**, vectors are bold-face small letters, e.g., **v**, while for scalars small letters in italics are used, e.g., *s*.

[2] Gabor components can be expressed in terms of amplitude and phase, $J_{km} = |J_{km}|e^{i\phi_{km}}$. We refer to the amplitudes $|J_{k\mu}$ as Gabor magnitudes.

nodes, and finds high similarity between the labels of the corresponding nodes. Neural implementations of LGM have previously illustrated its biological plausibility as a method for pattern recognition, exemplified in the Dynamic Link Architecture (DLA)[3] (von der Malsburg, 1981; von der Malsburg, 1988; Willshaw & von der Malsburg, 1976). In this paper, we will employ an algorithmic formulation of LGM.

In our implementation, a graph is a rigid lattice of Gabor Jets (Fig. 1). The traditional measure of similarity s (Lades, 1994; Wiskott et al., 1997) between two Gabor Jets $\mathbf{f}_i$ and $\mathbf{f}_j$ is the cosine of the angle between the two jets:

$$s(\mathbf{f}_i, \mathbf{f}_j) = \frac{\mathbf{f}_i^T \mathbf{f}_j}{\|\mathbf{f}_i\| \, \|\mathbf{f}_j\|} \tag{2}$$

This measure is useful as it provides robustness to the amount of contrast. The similarity between two graphs, G and G' with sets of node labels $V = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$ and $V' = \{f'_1, f'_2, \ldots, f'_m\}$, respectively, is

$$S(G, G') = \sum_{i=1}^{n} s(\mathbf{f}_i, \mathbf{f}'_{q(i)}) \tag{3}$$

where $q(i) = j$ is the index of the jet to which $\mathbf{f}_i$ has been mapped.

In our implementations, the spacing between neighboring graph nodes is seven pixels. Using 128×128 images, this leads to graphs containing 120–150 nodes, depending on the size of the scenes. This sparse sampling is only used for the stored models in the memory, called model graphs. The images of the scenes to be recognized are represented without spatial subsampling (i.e., Gabor jets are taken at every pixel). Even in the representation provided by the model graphs, however, the patterns are represented in their entirety due to the overlap in both space and frequency domains between receptive fields of neighboring Gabor wavelet kernels.

LGM is inherently capable of coping with variations in size, translation, rotation, deformation, etc. In this paper, however, we will restrict our investigations to stimuli (Section 4.1) with only translation within cluttered scenes, such that only the spatial coordinates need to be searched in the graph matching process. Extensions to other image variations can simply be implemented by extending the search to include these additional dimensions.

3.2. LGM$^+$: labeled graph matching enhanced with lateral excitation

In our new variant of LGM, we augmented the graph similarity function with an element that emphasizes the topological coherence of the match. The new graph similarity function $\tilde{S}$ involves the enhancement of each pairwise similarity value s by its neighboring similarity values.

$$\tilde{S}(G, G') = \sum_{i=1}^{n} \tilde{s}(\mathbf{f}_i, \mathbf{f}'_{q(i)}) \tag{4}$$

$$\tilde{s}(\mathbf{f}_i, \mathbf{f}'_{q(i)}) = s(\mathbf{f}_i, \mathbf{f}'_{q(i)}) + s(\mathbf{f}_i, \mathbf{f}'_{q(i)}) \sum_{r} s(\mathbf{f}_r, \mathbf{f}'_{q(r)}) \tag{5}$$

where r is the index of neighbors of $\mathbf{f}_i$ in graph topology, and $q(r)$ the index of the jet matched with $\mathbf{f}_r$. The data shown in the paper were obtained with a neighborhood function which spans two jets away from $\mathbf{f}_i$, however the neighborhood can be restricted to only the immediate neighbors (i.e., one jet away) and the performance would degrade only slightly. The biological analog of this function can be found in the lateral excitatory interactions among the neighboring V1 hypercolumns (Gilbert, 1992).

The second term in Eq. (1) is the excitation received by $s(\mathbf{f}_i, \mathbf{f}'_{q(i)})$. As it can be seen, the amount of excitation directly depends on the value of $s(\mathbf{f}_i, \mathbf{f}'_{q(i)})$. This is consistent with the physiological findings in the visual cortex. It has been shown that '…the level of excitation induced by activating the horizontal inputs depends on the level of depolarization of the target cell: the more depolarized the cell is, the larger the excitatory postsynaptic potential, as the result of voltage-dependent sodium conductances. Thus, one can think of the effect of the horizontal connections as being state dependent and influenced by the level of activation of other inputs converging onto the cell (Hirsch & Gilbert, 1991)' (Gilbert, 1992, pp. 125). The function of this lateral excitation during graph matching is discussed in detail in the Discussion Section. As will be demonstrated below, the small change of adding lateral excitation to the original LGM has a profound impact on the performance of LGM.

3.3. Mutual information maximization

Viola's method of matching two images (Viola, 1995; Viola & Wells, 1995) is based on the concept of mutual

[3] In this architecture two separate neural layers represent the input and the stored graphs. Graph nodes and graph links are represented by neurons and excitatory connections between the neurons, respectively. Node labels are represented by the receptive field profiles of the neurons. The correspondence between the two graphs is found through a dynamical process where at the final state each neuron in the input layer is linked (or mapped on) to its corresponding neuron in the stored model layer, and the global pattern of the links (implicitly) represents the type of global transformation the model has undergone. The matching process is driven primarily by transient or dynamic links between the neurons of the two layers. Typically a fast synaptic plasticity mechanism mediates this process. Within a layer, neighboring neurons, and between the layers those with similar labels (and neighbors, after the initial stage) are correlated. Dynamic links are established between neurons whose temporal patterns of activation are correlated. These links, in turn, induce temporal correlation between the neurons they link together. This positive feedback loop will stabilize when all the nodes that correspond to each other are linked together in a coherent one-to-one fashion. Therefore, temporal correlation underlies a binding between nodes which eventually serves as the basis for finding correspondences (for a more detailed account of the dynamics see (Konen & von der Malsburg, 1992; Lades et al., 1993; Willshaw & von der Malsburg, 1976)). The anatomical correlates of the input and model layers can be conceived to be V1 (and/or V2) and inferotemporal (IT) cortex, respectively. Dynamic links can be interpreted as synaptic weights modified in a Hebbian fashion.

information maximization (MIM), where the best correspondence of an image with a template is determined by the alignment that obtains the highest value of mutual information:

$$I(\mathbf{U},\mathbf{V}) = H(\mathbf{U}) + H(\mathbf{V}) - H(\mathbf{U},\mathbf{V}) \tag{6}$$

Here $H(\cdot)$ denotes the entropy (or joint entropy, in the last term), and $\mathbf{U}$ and $\mathbf{V}$ are sets of data points $\mathbf{U}_i$ and $\mathbf{v_i}$ (with $i = 1 \ \ldots \ n$), respectively, for the two images or image regions to be matched. The order of the data points in $\mathbf{U}$ and $\mathbf{V}$ is not arbitrary but rather according to the order of pixels in the images, e.g., in an image row-by-row fashion. Without this ordering, the joint entropy in Eq. (6) would be ill defined. The joint entropy requires pairing data points $\mathbf{u_i}$ and $\mathbf{v_j}$. This pairing is analogous to the pairing in jet correlation Eq. (2) in LGM, hence this ordering also introduces a sensitivity to topology in Eq. (6) comparable to the original LGM algorithm. Empirical values for the entropy measure can be obtained based on Parzen density estimates of the image probability densities (Duda & Hart, 1973):

$$H(\mathbf{W}) = n\,E\{-\ln p(\mathbf{W})\}$$

$$\approx -\sum_{\boldsymbol{\mu}\in\Im} \ln\left(\frac{1}{n-1}\sum_{\boldsymbol{\mu}'\neq\boldsymbol{\mu}\in\Im}\varphi(W_{\boldsymbol{\mu}} - W_{\boldsymbol{\mu}'}, \Sigma)\right) \tag{7}$$

$$\varphi(W_{\boldsymbol{\mu}} - W_{\boldsymbol{\mu}'}, \Sigma) =$$

$$\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{1}{2}(W_{\boldsymbol{\mu}} - W_{\boldsymbol{\mu}'})^T\Sigma^{-1}(W_{\boldsymbol{\mu}} - W_{\boldsymbol{\mu}'})\right)$$

The elements of the covariance matrix Σ of the Parzen window kernel can be optimized by leave-one-out cross validation of the Parzen density estimate (Viola & Wells, 1995). In the context of images, a critical step in the mutual information formulation lies in defining the probability density of an image. Viola simply used the grayscale values of all pixels as entries of $\mathbf{W}$. Thus, the density of an image was obtained from a one-dimensional Parzen window estimate where Σ has only one coefficient to be optimized, while the joint entropy in Eq. (6) requires a two-dimensional Parzen window with three coefficients to be optimized. There are two assumptions in this approach: (i) all pixels of an image have the same probability density and (ii) all pixels are independent of each other. Although it is well known that these assumptions are not true for real images, the algorithm seems to work well in practice.

In order to adapt Viola's method to the Gabor jet representation, the Gabor jets for each spatial coordinate replace the grayscale pixel values in $\mathbf{u}$ and $\mathbf{v}$. This change of representation introduces one new problem: Parzen windows now need to be applied to 12-dimensional data for the image densities, and 24-dimensional data for the joint densities, but Parzen windows usually do not work well for high dimensional data. However, as we confirmed empirically, the jets of an image actually form low dimensional clusters. Under these circumstances, Parzen windows still work properly and the 'curse of dimensionality' is circumvented (Scott, 1992). In optimizing Σ, for the image density of $\mathbf{U}$ and $\mathbf{V}$, we assumed that Σ is a diagonal matrix with equal coefficients on all diagonal coefficients:

$$\Sigma = \mathbf{I}\,\sigma^2 \tag{7}$$

where $\mathbf{I}$ denotes the identity matrix. Thus, only one coefficient needed to be optimized by cross validation, as in Viola's original method. For the joint density estimation, Σ was partitioned into four equal blocks:

$$\Sigma_{\text{joint}} = \begin{bmatrix} \Sigma_{\mathbf{uu}} & \Sigma_{\mathbf{uv}} \\ \Sigma_{\mathbf{vu}} & \Sigma_{\mathbf{vv}} \end{bmatrix} = \begin{bmatrix} \mathbf{I}\sigma^2_{\mathbf{uu}} & \mathbf{I}\sigma_{\mathbf{uv}} \\ \mathbf{I}\sigma_{\mathbf{vu}} & \mathbf{I}\sigma^2_{\mathbf{vv}} \end{bmatrix} \tag{9}$$

where each block was a diagonal matrix with equal diagonal coefficients, such that effectively only three coefficients, $\sigma^2_{\mathbf{uu}}$, $\sigma^2_{\mathbf{vv}}$, $\sigma_{\mathbf{uv}} = \sigma_{\mathbf{vu}}$ had to be optimized. The optimal coefficients in all density estimates were determined from an exhaustive search over a grid of reasonable values, constrained by the requirement of positive definiteness of Σ.

In our experiments described in Section 4.1, for the computation of the density estimates, at each candidate match, we use the same set of Gabor jets as those used by the graph matching search (i.e., jets falling on a regular grid). The search strategy is also identical across all three methods—LGM, LGM$^+$, and gabor-based MIM. The scene is uniformly and exhaustively searched in five pixel intervals. Thus, by using the exact same features (at every single step of the search process) the difference between the LGM, LGM$^+$, and MIM methods is narrowed down to the difference between the similarity functions in Eqs. (3), (4) and (6), respectively.

4. Empirical evaluations

4.1. Task and stimuli

The pattern recognition task we chose as the test bed of our evaluations is object detection, the task of finding an object within a scene. We selected this task because both LGM and MIM methods have successfully been applied to this problem. To ensure objectivity, we used a set of stimuli to which neither method has been previously applied. The stimuli were scenes composed of digital embryos (Brady, 1999; Brady, 1998). Digital embryos are 3-D structures generated by a stochastic process that is modeled after an embryological process (Brady, 1999). Two examples of these objects are shown in Fig. 2. As can be seen, digital embryos are highly irregular, and they provide an interesting test bed for the study of object vision, as they resemble plants and animals. It has also been shown (Brady, 1999; Brady, 1998) that human subjects can extract and recognize these patterns based on the types of scenes shown in Fig. 3(c) and (f).

We refer to the embryo to be detected as the model

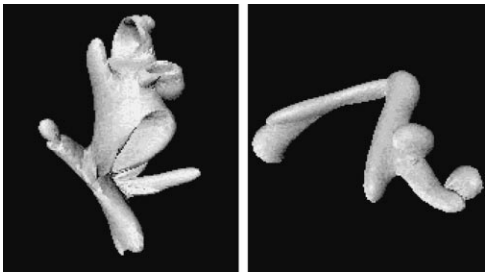

Fig. 2. Two examples of digital embryos.

embryo. The algorithms were tested using three arbitrary model embryos. Two of these embryos are shown in Figs. 3(a) and 4(a). For each model embryo we used 50 scenes containing the embryo, where half of the scenes contained occlusions. Six examples of the scenes corresponding to the first and second embryo are shown in Figs. 3(c) and 4(c), respectively. As can be seen in these pictures, in each scene, the model embryo is embedded in a background consisting of a clutter of various other embryos. Scenes were generated at random in terms of the choice and combination of the embryos constituting the background and the position of the model embryo within the scene. The complexity of the scenes is 'unbiased' in so far that they were generated by the inventor of the digital embryos without any knowledge about our research goals (prior to this study and for a different application). The irregular form of the embryos and the cluttered background consisting of the same type of shapes makes the 'foreground' (or model) embryo indistinct and non-segmentable from the background in the absence of top-down knowledge about the model. Although the embryo is three-dimensional, approximately the same viewpoint is used in all the scenes. In different scenes, because of translations of the model within the scene, minor variations in orientation in depth are present due to the change in the relative position of the object to the virtual camera (i.e., viewpoint of rendering of the image). Occluded scenes were created by superimposing a grid pattern over the cluttered scenes. Due to the translations of the embryo across the scenes, varying segments of the embryo are occluded in different scenes. These images were meant to mimic the occlusion caused by viewing scenes from behind a grid window.

In order to test the robustness of the methods with respect to a variation other than occlusion, we also obtained two images of the model embryos in Figs. 3(a) and 4(a) which differed from those embedded in the scenes in terms of lighting. The variants of these embryos are displayed in Figs. 3(b) and 4(b), respectively. Both MIM and LGM methods have been claimed to cope well with common image variations (Okada, Steffens, Maurer, Hong, Elagin,

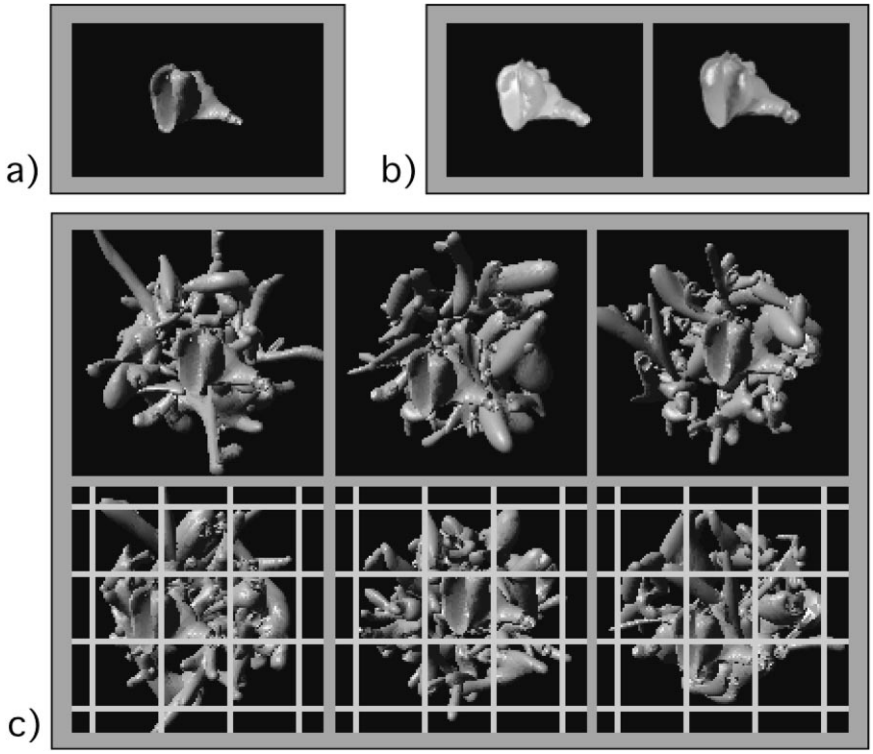

Fig. 3. Digital embryos and digital embryo scenes. a) An image of the first model embryo. b) Two other images of the embryo shown in (a) under different illuminations. c) Top row: three examples of the 25 scenes without occlusions against which the detection of embryos in images with (a) and (b) were tested. Bottom row: three examples of the scenes with occlusions. In each scene the model embryo is embedded in a random position in a background which is composed of a number of randomly chosen other embryos. In the no variation condition the image in (a) was used as the model to be detected in the scenes exemplified in (c). In the variation condition, the embryos shown in b) were used as models to be detected in the scenes exemplified in (c).

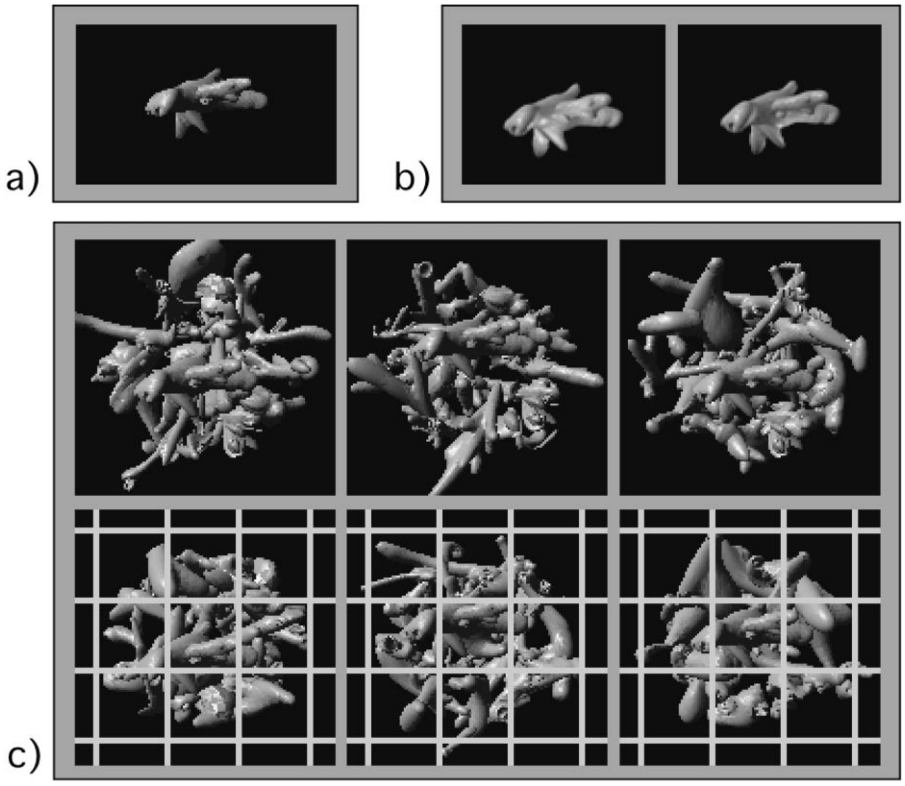

Fig. 4. Another digital embryo. See caption of Fig. 3 for explanation.

Neven, & ven der Malsburg et al., 1998; Viola & Wells, 1995). Our tests, therefore, involve four conditions: (1) the model embryo to be detected in the cluttered scenes is (almost) identical to the embryo image embedded in the scenes—the no variation condition (using three embryos each searched in 25 scenes), (2) the model embryo is substantially different in terms of lighting from those embedded in the scenes—the lighting variation condition (using four embryos each searched in 25 scenes), (3) the model embryo is detected in scenes where the embedded embryo is partially occluded—the occlusion variation condition (using three embryos each searched in 25 scenes) and (4) the model embryo is substantially different in terms of lighting from those embedded in the scenes and is to be detected in scenes where there is partial occlusion of the embryo—the lighting and occlusion condition (using four embryos each searched in 25 scenes).

Four algorithms were tested: MIM on grayscale pixel values (i.e., exactly as prescribed by Viola's original method (Viola, 1995; Viola & Wells, 1995)), MIM on the Gabor responses, the original LGM, and LGM^+. As a control we also tested pixel correlation.

4.2. Results

Before examining the performance of the four algorithms, we assessed the level of difficulty of the object detection task with a control experiment using the correlation method based on the grayscale pixel images. The gray-level correlation method resulted in perfect performance in the *no variation* condition, however, the performance dropped to 4 and 10% in the *lighting* and *lighting&occlusion* conditions. This test confirmed that the illumination variations presented a significant change in the gray-level distribution of the images, and the detection task in these conditions is not trivial anymore.

The performances of the four methods in the four conditions are illustrated in Fig. 5. The vertical axis denotes the percentage of the correct detection of the model embryo within the 25 scenes. The bars represent the average performance over the test patterns per condition, i.e., three different model embryos in the *no variation*, and *occlusion* conditions and four model embryos in the *lighting*, and *lighting&occlusion* conditions, and error bars denoting standard errors are provided in the plots as well.

The high performance of the MIM method with jets in the no variation condition indicates that an entropy method can operate well on high dimensional features such as Gabor jets. It also confirms that our strategy for optimizing the Parzen density estimation parameters is adequate. However, LGM^+, our new LGM method with lateral excitation achieve the best results throughout all tests.

To compare the relative performance of the methods, we performed pairwise two-tailed *t*-tests on the results pooled across all conditions, as summarized in Table 1. Comparison between performance of pixel and Gabor jet representations

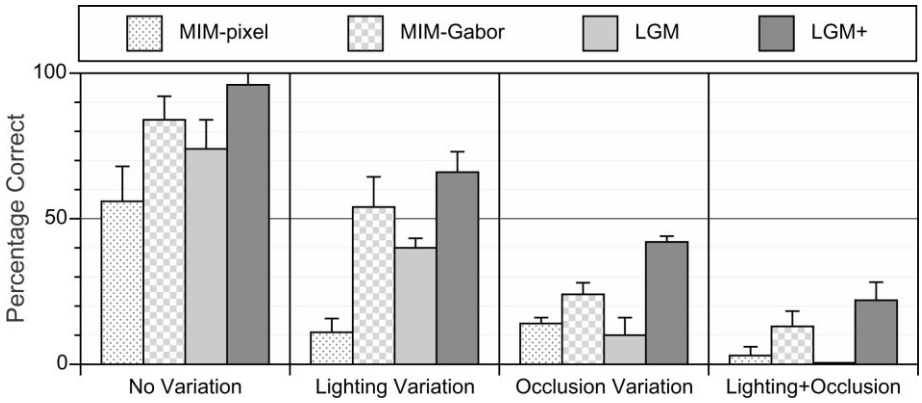

Fig. 5. Performance of the methods in four different conditions. The vertical axis represents the percentage of the correct detection in a set of 25 scenes. Each bar represents the average of the performance over the model embryos used in each condition (see text). 'MIM-pixel' and 'MIM-Gabor' refer to the MIM method operating on gray-level pixels, and operating upon Gabor jets, respectively. 'LGM' and 'LGM$^+$' represent the traditional Labeled Graph Matching and that augmented with lateral excitation, respectively.

for the MIM method clearly shows that the Gabor jet representation is significantly superior to the pixel representation. This is not surprising, as Gabor wavelets provide robustness to small changes in size, rotation, illumination, noise, etc. The performance of LGM is superior to that of pixel-based MIM, but not the Gabor-based MIM. Finally, the comparison of the methods when based on Gabor jets demonstrates that the performance of LGM$^+$ is significantly better than that of the traditional LGM and MIM.

5. Discussion

Labeled Graph Matching (LGM) (von der Malsburg, 1988) was introduced over a decade ago as a biologically inspired method for pattern recognition. Our results demonstrated that despite its age, LGM remains a competitive algorithm, even in comparison with the modern analytically developed statistical information processing methods. The original version of MIM, based on gray-scale pixel values, had the worst performance. However, replacing the gray-level features with Gabor jets as the basic representation improved it so significantly that it performed superior to LGM. Incorporating a biologically plausible lateral excitation mechanism in LGM, on the other hand, improved its performance to the extent that LGM$^+$ significantly outperformed the Gabor-based MIM.

Several reasons may account for the superiority of LGM$^+$ to (Gabor-based) MIM. First, LGM$^+$ incorporates topological constraints that are not available to the MIM algorithm; it is not obvious how to employ such additional constraints efficiently in MIM. Second, the MIM algorithm assumes that individual pixels in the image are independent and have the same probability distribution of the image features—an assumption that is usually not correct. Third, the MIM algorithm requires careful parameter optimization for the Parzen density estimate upon which mutual information computation is based. As the optimal values of these parameters need to be found in a landscape which has usually several local minima, the parameter search can be brittle and get stuck in suboptimal values (Viola, 1995).

An interesting question concerns the characteristic of LGM$^+$ that underlies its superiority over LGM (and the other methods examined here). Examining Eq. (5) reveals the following characteristics. Given equal similarity values across all the nodes in the graph for a given match, the nodes at the graph boundary receive less excitation from their neighbors than the nodes in the center, because they have fewer neighbors. In other words, given random similarity values, the nodes at the object boundary are weighted less than those in the center, and thus their contribution to the total graph similarity is reduced. Of course, this argument only holds on average; in cases where a boundary node is surrounded by very high similarity values, or where a center node is surrounded by low similarity nodes, this will no longer be the case. However, this subtle weighting scheme can be important in situations where the objects are to be detected in cluttered backgrounds[4], as in our scenes. In such

Table 1
Statistical significance of relative performance between pairs of the four tested algorithms, using data pooled across all experimental conditions for each method. The values show that each of the methods performed significantly different from each of the other methods. The differences between the means within each condition were also significant but not shown in the table. The ranking (in decreasing order) of the methods, as can be seen is: (1) LGM$^+$, (2) MIM-Gabor, (3) LGM and (4) MIM-pixel

t-test values	MIM-Gabor	LGM	LGM$^+$
MIM-pixel	$P < 0.005$	$P < 0.05$	$P < 0.0005$
MIM-Gabor		$P < 0.005$	$P < 0.0005$
LGM			$P < 0.0001$

[4] For a discussion of the effects of background on pattern matching, and possible remedies please see Pötzsch, Krüger, & von der Malsburg (1996); Würtz (1995), Würtz (1997). A functional approach to weighting graph nodes is discussed in Krüger (1997).

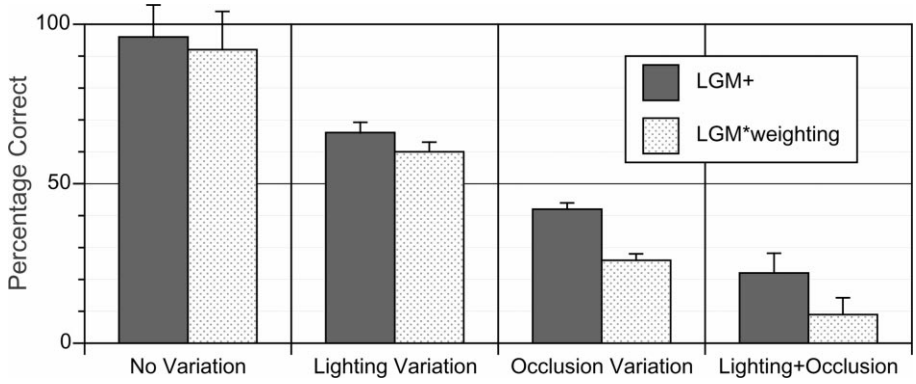

Fig. 6. Comparison of LGM$^+$ with and LGM algorithm that uses fixed nearest-neighbor node weighting.

scenes, the jets located at or near the boundary of the object are affected by the structure in the background, hence leading to low similarity with the boundary jets in the model graph. Lateral excitation in effect weighs down the contribution from the boundary jets and improves the robustness of graph matching.

To investigate whether the success of LGM$^+$ is due merely to this node weighting we compared its performance with an analogous algorithm which encodes weighting of the nodes explicitly based on their number of neighbors, and hence proximity to the boundary. The algorithm was implemented by substituting Eq. (5) with:

$$\tilde{s}(\mathbf{f}_i, \mathbf{f}'_{q(i)}) = s(\mathbf{f}_i, \mathbf{f}'_{q(i)}) + s(\mathbf{f}_i, \mathbf{f}'_{q(i)}) \sum_r 0.75 \tag{10}$$

The second term reflects the fact that the similarity between two random jets is on average 0.75, a value determined from empirical evalutations. Fig. 6 displays the comparison between the performance of the LGM algorithm enhanced with lateral excitation of Eq. (5) (i.e., the LGM$^+$ method) and the LGM algorithm employing the weighting of Eq. (10). Although the performances of the two methods are comparable on scenes with no occlusion—where the main difficulty is the background effect—the lateral excitation method outperforms weighting systematically in the occlusion conditions ($p < 0.005$). This differential performance on the two conditions demonstrates that different characteristics subserve the strengths of the two methods, and that the strength of LGM$^+$ is not merely due to weighting. While weighting is fixed across matches, lateral excitation between the nodes is dynamic and changes from one match to another. Over the course of the matching process, lateral excitation favors matches with contiguous (or topographically smooth) high similarity profiles over matches that contain topographically isolated high similarity values (non-smooth high-value profiles). This, in turn, favors correct graph correspondences over the incorrect ones. The false correspondences which may have equal or higher total graph similarity than the correct correspondence, tend to be based on a few accidental and hence topologically sporadic high jet similarities, whereas correct graph correspondences tend to have graph similarities which are much more coherent both in terms of their value and topology. It is this coherence that is rewarded by LGM$^+$'s lateral excitation scheme. The addition of variations to the pattern detection task, in particular occlusions, decreases the total similarity of the correct correspondence and consequently introduces an increasing number of false matches. For this reason, the superiority of LGM$^+$ becomes more pronounced with increasing amount of difficulty in the matching task, as nicely illustrated in Fig. 6. The strength of lateral excitation (and its superiority to LGM) has also been shown and discussed in a previous study where the stimuli had *no* background and matching had to cope with partial information as well as variations in size and proportions (Shams, 1999). In such no-background stimuli, weighting of boundary nodes cannot explain the improvements shown by LGM$^+$

6. Conclusion

We compared two Labeled Graph Matching algorithms with a state-of-the-art information theoretic algorithm—mutual information maximization—in the task of object detection. The original version of MIM had the worst performance since it operated out of gray-level pixels, a representation that is known to be brittle in face of variations of lighting, object size, occlusions, noise, etc. Our modified version of this algorithm replaced the pixel values with a Gabor wavelet representation and improved it so significantly that it performed superior to a highly successful version of Labeled Graph Matching. However, our new LGM$^+$ algorithm that employs lateral interactions in the graph could achieve significantly better object detection results than MIM. This seems to be primarily due to the fact that LGM$^+$ incorporates topological constraints that are not available to the MIM algorithm.

The need for parameter search makes the MIM algorithm also significantly more computationally expensive: even with our highly optimized code for density estimation, MIM required 2–3 orders-of-magnitude more floating point operations than LGM/LGM$^+$. In contrast, LGM$^+$, as

the original version, is computationally simple and does not require any on-line parameter optimization.

7. List of mathematical symbols

$\Im$	Two dimensional lattice of pixel positions
$\boldsymbol{\mu}$	Pixel position
$\mathbf{k}$	Vector determining the orientation and center spatial frequency of the wavelet transform
W	Gray-scale image
$\psi_{\mathbf{k}}(\boldsymbol{\mu})$	Gabor wavelet kernel centered oriented along vector $\mathbf{k}$ and centered around spatial coordinate $\boldsymbol{\mu}$
$J_{\mathbf{k}\boldsymbol{\mu}}$	Gabor wavelet transform of an image with kernel $\psi_{\mathbf{k}}(\boldsymbol{\mu})$
$\mathbf{f}_i$	Vector of Gabor wavelet transforms $J_{\mathbf{k}\boldsymbol{\mu}}$ for a given $\boldsymbol{\mu}$ but varying $\mathbf{k}$; referred to as a Gabor Jet
$s(\mathbf{f}_i, \mathbf{f}_j)$	Similarity between Gabor jets $\mathbf{f}_i$ and $\mathbf{f}_j$
G	Graph
V	Set of a graph vertices or Gabor jets
$q(i)$	Correspondence function that maps jet $\mathbf{f}_i$ of graph G to jet of graph G'
$S(G,G')$	Similarity between graphs G and G' used by LGM method
$\tilde{S}(G,G')$	Similarity between graphs G and G' used by LGM$^+$ method
$H(W)$	Entropy of image W
$I(\mathbf{U}, \mathbf{V})$	Mutual information between images U and V
Σ	Covariance matrix
$\lvert\Sigma\rvert$	determinant of Σ
n	Total number of pixels in image W
$p(\mathbf{W})$	Probability density of image W
$\mathbf{I}$	Identity matrix

Acknowledgements

We are grateful for the comments of the anonymous reviewers, in particular for helping us improve the discussion with respect to the difference between LGM and LGM$^+$. This work was supported by NSF's IMSC program at USC, by NSF award no. 9710312, the ERATO Kawato Dynamic Brain Project, the ATR Human Information Processing Research Laboratories, and NIH grant HD08506.

References

Becker, S. (1995). JPMAX: Learning to recognize moving objects as a model-fitting problem. *Advances in Neural Information Processing Systems*, *7*, 1–9.

Bienenstock, E., Geman, S., & Potter, D. (1997). Compositionality, MDL priors, and object recognition. In M. C. Mozer, M. I. Jordan & T. Petsche, *Advances in neural information processing systems* (pp. 838–844). Vol. 9. MIT Press.

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

Brady, M. (1999). Psychophysical investigations of incomplete forms and forms with background. Unpublished PhD Thesis, University of Minnesota.

Brady, M. J. (1998). Learning to recognize camouflaged novel objects. Paper presented at the The Association for Research in Vision and Ophthalmology Meeting, Florida.

Buhmann, J., Lange, J. & von der Malsburg, C. (1989). Distortion invariant object recognition by matching hierarchically labeled graphs. Paper presented at the IJCNN International Conference on Neural Networks, Washington, DC.

Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P. & Marchal, G. (1995). Automated multimodality image registration using information theory. Paper presented at the Information Processing in Medical Imaging, Ile de Berder, France.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*, New York: Wiley.

Gilbert, C. D. (1992). Horizontal integration and cortical dynamics. *Neuron*, *9* (1-13), 121–128.

Grossmann, A., & Morlet, J. (1985). *Decomposition of functions into wavelets of constant shape, and related transforms, mathematics and physics, lecture on recent results*, Singapore: World Scientific Publishing.

Hirsch, J. A., & Gilbert, C. D. (1991). Synaptic physiology of horizontal connections in the cat's visual cortex. *Journal of Neuroscience*, *11*, 1800–1809.

Konen, W., & von der Malsburg, C. (1992). Unsupervised symmetry detection: a network that learns from single examples. In I. Aleksander & J. Taylor, *Artificial Neural Networks*, Elsevier.

Konen, W., & Vorbrüggen, J. C. (1993). *Applying dynamic link matching to object recognition in real world images*, FRG: Institut für Neuroinformatik, Ruhr-Universität Bochum.

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

Krüger, N. (1997). An algorithm for the learning of weights in discrimination functions using a priori constraints. *IEEE transactions on pattern analysis and machine intelligence*, *19* (7), 764–768.

Krüger, N., Pötzsch, M., & von der Malsburg, C. (1997). Determination of face position and pose with a learned representation based on labeled graphs. *Image and Vision Computing*, *15*, 665–673.

Lades, M. (1994). Invariant object recognition with dynamical links, robust to variations in illumination. Unpublished PhD Thesis, Ruhr-Universität Bochum.

Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange v.d., J., Malsburg, C., Würtz, R. P., & Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transaction on Computers*, *42* (3), 300–311.

Maurer, T. & von der Malsburg, C. (1995, October 9–13). Learning feature transformations to recognize faces rotated in depth. Paper presented at the International Conference on Artificial Neural Networks, Paris, France.

Maurer, T. & von der Malsburg, C. (1996). Tracking and learning graphs and pose on image sequences of faces. paper presented at the Proceedings of the 2nd International Conference on Automatic Face- and Gesture-Recognition, Killington, Vermont, USA.

McGarry, D. P., Jackson, T. R., Plantec, M. B., Kassell, N. F. & Downs, J. H. (1997). Registration of functional magnetic resonance imagery using mutual information. Paper presented at the SPIE Medical Imaging.

Moskalik, A., Carson, P. L., Meyer, C. R., Fowlkes, J. B., Rubin, J. M., & Rubidoux, M. A. (1995). Registration of 3-D compound ultrasound scans of the breast for refraction and motion correction. *Ultrasound in Medicine and Biology*, *21* (6), 769–778.

Okada, K., Steffens, J., Maurer, T., Hong, H., Elagin, E., Neven, H., & ven der Malsburg, C. (1998). The Bochum/USC face recognition system and how it fared in the FERET phase III test. In H. Wechsler, P. Phillips, V. Bruce, F. Soulie & T. Huang, *Face recognition: from theory to applications*, Springer-Verlag.

Pluim, J. P. W., Maintz, J. B. A., & Viergever, M. A. (2000). Image registration by maximization of combined mutual information and

gradient information. *IEEE Transactions on Medical Imaging*, *19* (8), 809–814.

Pötzsch, M., Krüger, N., & von der Malsburg, C. (1996). Improving object recognition by transforming gabor filter responses. *Network: Computation in Neural Systems*, *7* (2), 341–347.

Rowley, H. A., Baluja, S. & Kanade, T. (1995). Human face detection in visual scenes (Internal Report CMU-CS-95-158): Carnegie Mellon University, School of Computer Science.

Scott, D. W. (1992). *Multivariate density estimation*, Wiley: New York.

Shams, L. (1999). Development of visual shape primitives. Unpublished PhD Thesis, University of Southern California, Los Angeles.

Shams, L., & von der Malsburg, C. (1999). Are Object Shape Primitives Learnable? *Neurocomputing*, *26-27*, 855–863.

Studholme, C., Hill, D. & Hawkes, D. J. (1996, Sept. 1996). Automated 3-D registration of truncated MR and CT images of the head. Paper presented at the Brit. Mach. Vis. Conf.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Neuroscience*, *3* (1), 71–86.

Viola, P. (1995). Alignment by maximization of mutual information. Unpublished PhD Thesis, MIT, Cambridge.

Viola, P. & Wells, W. (1995). Alignment by maximization of mutual information. Paper presented at the Fifth International Conference on Computer Vision, Cambridge, MA.

von der Malsburg, C. (1981). The correlation theory of brain function (Internal Report 81-2): Dept. Neurobiology, Max-Planck-Institute for Biophysical Chemistry, P.O. Box 2841, Göttingen, Germany.

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

von der Malsburg, C., Shams, L. & Eysel, U. (1998, November 1998). Recognition of images from complex cell responses. Paper presented at the Society for Neuroscience Meeting, Los Angeles, CA.

Willshaw, D. J., & von der Malsburg, C. (1976). How patterned neural connection can be set up by self-organization. *Pflugers Arch. Suppl.*, *359*, 463–469.

Wiskott, L. (1996). *Labeled graphs and dynamic link matching for face recognition and scene analysis (vol. 53)*, Frankfurt: Verlag Harri Deutsch.

Wiskott, L. (1996). *Recognizing objects in cluttered scenes, labeled graphs and dynamic link matching for face recognition and scene analysis* (pp. 73–80). , Vol. 53. Frankfurt: Verlag Harri Deutsch.

Wiskott, L., Fellous, J.-M., Krüger, N. & von der Malsburg, C. (1995, June 26–28). Face recognition and gender determination. Paper presented at the International Workshop on Automatic Face- and Gesture-Recognition, Zürich.

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

Wiskott, L., & von der Malsburg, C. (1993). A neural system for the recognition of partially occluded objects in cluttered scenes. *Int. J. Pattern Recognition and Artificial Intelligence*, *7* (4), 935–948.

Wiskott, L. & von der Malsburg, C. (1995). Recognizing faces by dynamic link matching. Paper presented at the International Conference on Artificial Neural Networks, Paris, France.

Würtz, R. (1995). *Multilayer dynamic link networks for establishing image point correspondences and visual object recognition*, Frankfurt am Main: Verlag Harri Deutsch.

Würtz, R. P. (1997). Object recognition robust under translations, deformations and changes in background. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19* (7), 769–775.