

Size and Distortion Invariant Object Recognition

by Hierarchical Graph Matching‡

JOACHIM BUHMANN*, MARTIN LADES* and CHRISTOPH VON DER MALSBURG*†

Computer Science Department,
Program in Neural, Informational and Behavioral Sciences†,
University of Southern California, University Park,
Los Angeles, Ca 90089-0782*

We present a neural system for invariant object recognition. Its flexibility is demonstrated with freely taken camera images of human faces. The system is an application of the Dynamic Link Architecture, which owes its strength to an enhancement of traditional neural networks by a new kind of variables to express hierarchical grouping of neurons. This capability is used here to group primitive local feature detectors (Gabor-based wavelets) into composite feature detectors (jets), and to preserve neighborhood relationships between jets when they loose position information on the way from the image domain to the object domain. Due to the potential for grouping, objects can be represented as attributed graphs, jets serving as attributes. Recognition is formulated as graph matching and is implemented as a topologically constrained diffusion of image-object links. We proceed here in a hierarchical sequence of matches, from low frequency components of jets to high frequency components. Size invariance is achieved by interposing diffusion steps in magnification space. The system is implemented on a network of transputers.

1. Introduction.

Some of the great promises of neural systems are ease of massively parallel implementation and ability to learn from examples. These capabilities have been amply illustrated in small systems, but in order to live up to expectations neural systems have to be scaled up to large systems. A serious impediment to scaling, at least with current learning mechanisms, is diverging learning time. The core of the problem is that discrimination of new pattern types is possible only with the help of newly trained specific feature types, which can be defined only after a full ensemble of patterns has been examined by the system.

The system presented here is based on the Dynamic Link architecture [1, 2, 3]. This neural architecture enhances traditional neural networks with the capability to form and handle groups of neurons and to assemble sets of neurons into structured graphs. Furthermore, inherent in the architecture is the capability to efficiently perform attributed graph matching [1, 4, 5]. On the basis of these capabilities it is possible to perform pattern discrimination with the help of an object-independent standard set of feature detectors, to automatically generalize over large groups of symmetry operations, and to acquire new objects by one-shot learning. As a consequence, all time-consuming learning steps are obviated for many important applications, and the door to large-scale systems is opened.

‡Supported by grants from the German Ministry for Science and Technology (ITR-8800-H1) and from the AFOSR (88-0274).

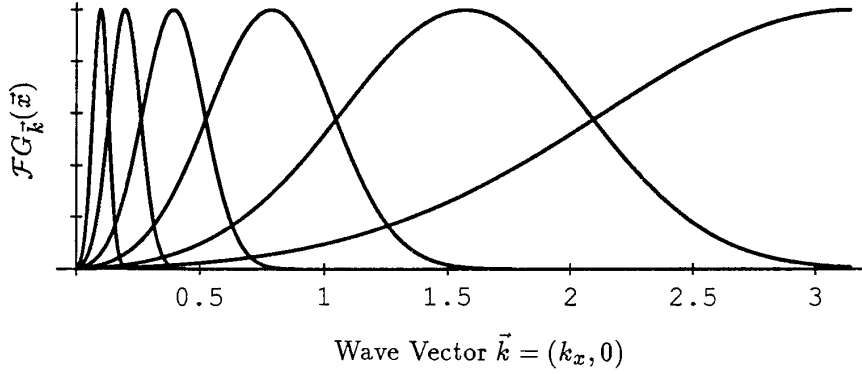


Figure 1: Coverage of frequency space by wavelets of different resolution levels.

The basic form of the system we are using here has been described elsewhere [6, 7]. It demonstrates invariant recognition of complex objects (human faces). It successfully deals with the most important invariances, translation within the image plane, perspective distortion, scaling, background variation, and moderate changes in illumination. The system optimizes ease of implementation in conventional processors and does not pay attention to neural style in details. Main advantages of neural systems are, however, preserved, among them the potential for massively parallel implementation, fault tolerance, and simplicity of formulation. A completely neural version has been described in [5].

Compared to the earlier versions [6, 7] of our object recognition system we have extended the graph matching procedure into a hierarchical optimization process. Starting with low frequency information, the system gradually uses finer and finer image detail for graph matching, and finally ends with a reliable classification of the object. A second improvement to the system is size invariance and size estimation.

2. Gabor based wavelets as a suitable data format for object recognition

Our first step towards achieving object recognition that is invariant with respect to background, translation, distortion and size is to choose a set of primitive features which is maximally robust with respect to such variations. We chose Gabor based wavelets. According to their definition, wavelets form a family of functions which are localized in space and in frequency space and all of which are scaled, rotated and translated versions of a single function. To be “admissible” their spatial integral has to exist. The wavelets we employ are based on the Gabor function and may be written as:

$$G_{\vec{k}}(\vec{x}) = \left(\exp(i\vec{k} \cdot \vec{x}) - \frac{k}{\sqrt{\pi}\sigma} \exp\left(-\frac{\sigma^2}{2}\right) \right) \exp\left(-\frac{\vec{k} \cdot \vec{k} \vec{x} \cdot \vec{x}}{2\sigma^2}\right). \quad (1)$$

The real and imaginary parts of G are spatially decaying sine and cosine waves, respectively. The wave vector \vec{k} of length $k \equiv \|\vec{k}\|$ defines the spatial wavelength and at the same time controls the width of the Gaussian window G , which is σ/k . We chose $\sigma = \pi$ to approximate the shape of receptive fields found neurophysiologically in visual cortex [8]. With this value, the real part of G has a central peak and two side lobes. The variation of the receptive field size is controlled by \vec{k} . The constant term subtracted from the oscillatory part in (1) ensures that the overlap of the function with a constant function vanishes and avoids dependence of the filter responses on global changes of illumination. We are working with quadratic images of size $N = 128$. Our entire wavelet family consists of 6 frequency bands, which are spaced in octaves, the wave vector \vec{k} assuming the length values

$$k_{\kappa} = \frac{2\pi}{N} 2^{\kappa} \quad \text{with } \kappa \in \{1, \dots, 6\}. \quad (2)$$

Accordingly, we have a logarithmic coverage of Fourier space, as shown in Fig. 1. Each wavelet with a given $\|\vec{k}\|$ comes in 8 different orientations. These are equally spaced to cover all orientations between zero and π , i.e., $\vec{k} = k(\cos \phi, \sin \phi)$ with $\phi = \pi\nu/8$, $\nu \in \{0, \dots, 7\}$.

Wavelets, characterized by their frequency, position and orientation, might be interpreted as feature detectors. To determine the degree of excitation of feature types at location \vec{x}_0 , the image is convolved with the subset of wavelet functions located at \vec{x}_0 :

$$(\mathcal{W}I)(\vec{k}, \vec{x}_0) := \int G_{\vec{k}}(\vec{x}_0 - \vec{x}) I(\vec{x}) d^2x, \quad (3)$$

where $I(\vec{x})$ is the image to be processed. In (3) we define the wavelet transform as a linear operator on the space of all images $I(\vec{x})$. We will now drop the property of linearity by introducing two nonlinear transformations of the wavelet coefficients. Firstly, we take the modulus of the transform, which varies smoothly across an intensity variation and will thus greatly alleviate our matching process. Secondly, within each frequency band we normalize the energy over the entire image to a constant. This lets receptive fields of all sizes contribute equally to the matching procedure. In many grey level images the high frequency coefficients are much smaller than those for low frequency, although they encode important structure. Our final formula for the set of features extracted from the image I at position \vec{x}_0 is:

$$\mathcal{J}I(\vec{k}, \vec{x}_0) = \frac{|\mathcal{W}I|}{\int |\mathcal{W}I| d^2x d\phi}. \quad (4)$$

(ϕ is the orientation of \vec{k} in the frequency plane.) We will refer to the set

$$\left\{ \mathcal{J}I(\vec{k}, \vec{x}_0) | \kappa \in \{1, \dots, 6\}, \nu \in \{0, \dots, 7\} \right\}$$

as the *jet* at position \vec{x}_0 . Although we discard information in (4), jets still contain enough structure to permit object identification.

3. Attributed Graph Matching and Size Estimation by Hierarchical Search

In our object recognition system, patterns are encoded as planar graphs (V, E) , with vertex set V and edge (link) set E . Neighboring vertices are connected by links, which encode information about the local topology. Vertices refer to locations in patterns, carry jets as attributes, and thus form local descriptors of object structure.

The complete object recognition system is composed of two modules, the *image domain* (I) and the *object domain* (O). Gray level images of size 128×128 pixels with 8 bit depth are preprocessed in the image domain by wavelet transformation [3] and ensuing jet formation [4]. We form the object domain by storing camera-derived portrait images (this is our “one-shot learning”). Each stored object in this gallery is formed by picking a rectangular grid of points as graph nodes, the grid being appropriately positioned over the face to be stored, and by storing with each grid point the jet locally determined. The face prototypes thus formed serve as pattern classes for our recognition system.

For recognition, a new image of the person to be recognized is taken, the image is transformed into jets (in the present system somewhat redundantly by computing jets for all pixel positions), and all stored object graphs are tentatively matched to the image. Conforming with the Dynamical Link Architecture [1, 2, 9, 6], this is done by establishing and dynamically modifying links between vertices in the object domain and vertices in the image domain, as explained below. This process amounts to a rapid change of connectivity within the network, a radical deviation from traditional neural networks, in which connectivity is essentially unchanged during individual processes. In the specific implementation we are reporting on here, this graph matching is formulated as an algorithm with the following steps.

Initialization: An object is selected from the object domain. A copy of the object graph is positioned in some central position in the image domain (“image graph”). Each vertex in the object graph is connected to the corresponding vertex in the image graph. A size parameter α , encoding relative size of the pattern in the image domain, is initialized to the value 1.

Evaluation of match quality: This is done with the help of the cost function

$$C_{\text{total}} := \lambda C_{\text{top}} + C_{\text{jet}}. \quad (5)$$

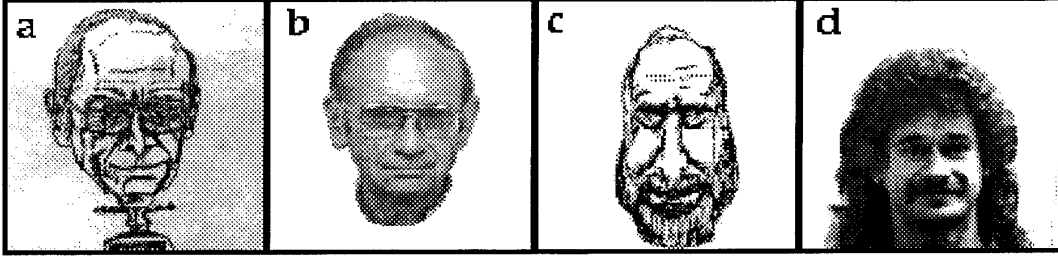


Figure 2: Example of face recognition. a) Image. b) Best matching object, out of a gallery of 17 faces. c) and d) Second and third best match.

The first term has the form:

$$C_{\text{top}} := \sum_{(i,j) \in E} \left(\frac{\tilde{\Delta}_{(i,j)}^I}{\alpha} - \tilde{\Delta}_{(i,j)}^O \right)^2. \quad (6)$$

In this “topology term,” $\tilde{\Delta}_{(i,j)}^I$ is the geometrical distance vector between vertices i and j in the image, $\tilde{\Delta}_{(i,j)}^O$ the same for the corresponding vertices in the object. This term penalizes deformations of the image graph compared to the object graph. The parameter λ gauges the relative importance of the topology term in (5) to the other one, the “similarity term:”

$$C_{\text{jet}} := - \sum_{i \in V} S_{\kappa} \left(\mathcal{J}_i^O, \mathcal{J}_i^I(\alpha) \right). \quad (7)$$

This is a sum of all similarities between pairs of corresponding jets, one taken from the image domain and one from the object domain. The function S_{κ} measures the similarity between two jets, where the index κ signifies that jet components are taken into account only from frequency level 1 up to and including level κ . Experimentation with different similarity functions [6, 7] showed that recognition results are not sensitive to the details of S . The simulations described here have been carried out with $S_{\kappa}(\mathcal{J}_i^O, \mathcal{J}_i^I) = \cos(\Theta)$, where Θ is the angle between the κ -truncated jets $\mathcal{J}_i^O, \mathcal{J}_i^I$.

Diffusion on level κ : One vertex after another is tentatively shifted to a random new position, which is accepted if C_{total} is decreased. Similarity of two jets is evaluated only on the basis of the jet components corresponding to the current κ . This update process continues until the graph has relaxed to a stable configuration. Diffusion guided only by low frequency information is equivalent to a sort of blob detection in the image. The first diffusion level, with $\kappa = 1$, also serves to position the image graph appropriately.

Jet transformation: The jet \mathcal{J}_i^I has to be transformed according to the size estimate α . The necessary transformation of the jet components for a scaled image $I'(x) = I(\alpha x)$ is given by the scaling relationship of the wavelet coefficients

$$(\mathcal{W}I')(k, x_0) = \frac{1}{\alpha} (\mathcal{W}I)\left(\frac{k}{\alpha}, \alpha x_0\right). \quad (8)$$

For scaling factors α such that $\frac{k}{\alpha}$ does not lie on the sampling grid $(k_{\kappa}, \kappa \in \{1, \dots, 6\})$ in Fourier space, jet components are linearly interpolated.

Size estimation: The image graph is scaled by a factor α' , keeping its center fixed. If C_{total} is reduced, the new value is accepted: $\alpha' \rightarrow \alpha$. The procedure is repeated until an optimum is reached.

Iteration. Diffusion and size estimation are repeated for increased resolution level until $\kappa = 6$ is reached. With each iteration more image structure is taken into account.

Object recognition: This sequence of steps is performed for each stored object separately and the optimal C_{total} is thus determined. In this way the object with the best match to the image is determined. If this quality is below a certain threshold, the image is classified as unknown to the system. If one object matches significantly better than all competitor objects, the face in the image is considered as recognized.

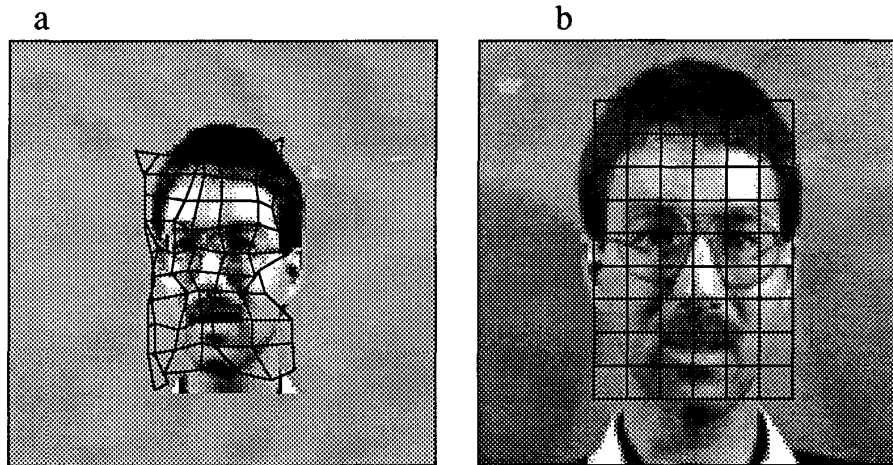


Figure3: Demonstration of size invariance. a) Stored object with object graph. b) Image with optimal image graph.

4. Discussion

The system identifies a person's face by comparing an extracted graph with a set of stored graphs. Accordingly, there are two steps: (i) a picture gallery is formed by storing portrait images as graphs in the object domain; (ii) a person is recognized by taking a picture and comparing the labeled graph extracted from this picture with all stored graphs in the gallery. The picture to be recognized may differ from the prototype face in facial expression, viewing direction, background, lighting condition, and size, within reasonable limits.

Our gallery contains at present over 40 different face images. We have made little effort to standardize conditions of picture taking, for instance, background and illumination vary considerably from image to image. In spite of this, recognition success of the system already in the present form is remarkably consistent. An extreme recognition result is demonstrated in Fig.2. The line drawing (a) has been compared with all images in the gallery. On the basis of the final matching costs the recognition system assigned the unknown image (a) to the class (b). The second and the third best matches are shown in (c) and (d). The high rank of (c) is due to it also being a line drawing, which strongly affects the high frequency components of jets.

Capability for size invariant recognition is illustrated in Fig.3. An image taken with smaller magnification (a) is compared with an object (b) stored in the gallery. Notice that the faces also differ in perspective. The rectangular grid in (b) shows the graph which encodes the stored image. At the end of the diffusion process the system found a shrunk graph which matches corresponding positions in the image fairly well. The final cost value was low enough to identify the image in (a) with the face in (b).

There are many obvious directions in which our system can be improved. One is the introduction of hierarchical structure in the picture gallery such that an early, low resolution, stage of our hierarchical process can already select a small sector of the object domain for further examination, reducing the number of required matches. Another is the introduction of more global parameters (analogous to our scaling factor α), to allow for even greater flexibility. Obvious choices are orientation and illumination. Already in its present form our system could easily be applied to object types other than faces and even to other perceptual domains. There is nothing to specialize it to the applications we have illustrated.

The main purpose of our system is to show that a neural system gains enormously in power if provided with a mechanism for grouping. In our application, this is important to bundle elementary feature detectors (wavelet components) into local composite features (jets), to express neighborhood relationships within the image in order to preserve topology while generalizing over position, and to bind image points to object points. The binding capability frees our system from the necessity to go through extensive learning cycles, thus removing the main barrier to the application of neural networks on a large scale. Although our specific version does not emphasize neural style in detail, it displays essential neural features, among them learning from examples and ease of parallel implementation. The latter is demonstrated by our implementation in a

network of transputers [7].

Acknowledgements: The Transputer system used here is a modified version of the one developed by Jan Vorbrüggen and Rolf Würtz, whose contribution we gratefully acknowledge. We thank Alfredo Weitzenfeld for permission to reproduce the caricatures in fig. 2.

- [1] C. von der Malsburg, "The Correlation Theory of Brain Function," Int. Report 81-2, Dep. Neurobiol. MPI Biophys. Chemie (1981).
- [2] C. von der Malsburg, "Nervous Structures With Dynamical Links," Ber. Bunsenges. Phys. Chem. 89, 703-710 (1985).
- [3] C. von der Malsburg, "Am I Thinking Assemblies?," in: Proceedings of the Trieste Meeting on Brain Theory, October 1984. G.Palm and A.Aertsen, eds. Springer: Berlin Heidelberg, pp 161-176 (1986).
- [4] E. Bienenstock, C. von der Malsburg "A Neural Network for Invariant Pattern Recognition," Euophys. Lett. 4, 121-126 (1987).
- [5] C. von der Malsburg, "Pattern Recognition by Labeled Graph Matching", Neural Networks 1, 141-148 (1988)
- [6] J. Buhmann, J. Lange, C. von der Malsburg, "Distortion Invariant Object Recognition by Matching Hierarchically Labeled Graphs", Proceedings of IJCNN'89, Vol. I 155-159, Washington (1989).
- [7] J. Buhmann, J. Lange, C. von der Malsburg, J.C. Vorbrüggen, R.P. Würtz, "Object Recognition in the Dynamic Link Architecture — Parallel Implementation on a Transputer Network —", in: Neural Networks: A Dynamical Systems Approach to Machine Intelligence, B. Kosko (Ed.), (Prentice Hall, New York 1990) in press
- [8] J.P. Jones, L.A. Palmer, "An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex", J. Neurophys. 58, 1233-1258 (1987).
- [9] C. von der Malsburg, E. Bienenstock, "A Neural Network for the Retrieval of Superimposed Connection Patterns," Euophys Let 3, 1243-1249 (1986).