Distortion Invariant Object Recognition by Matching Hierarchically Labeled Graphs

J. Buhmann, J. Lange and C. von der Malsburg

Computer Science Department, University of Southern California, University Park, Los Angeles, Ca 90089-0782

A graph matching process for object recognition is proposed. It is applied to face recognition. Grey level images are represented by a resolution hierarchy of local Gabor components, which are all scaled and rotated versions of each other. The components centered on one image point form a "Gabor jet." A single jet provides a distortion insensitive local representation of part of an image. Object recognition is achieved by matching jets corresponding to a preselected set of image points to jets in stored prototype patterns. For a selected image jet the best match in each stored image is determined, under a constraint preserving the spatial arrangement of all jets. The procedure amounts to labeled graph matching, Gabor jets forming labels to nodes and topology determining links. A contrast insensitive similarity measure between image jets and stored jets provides for invariance with respect to lighting conditions. We have formulated the matching procedure as an optimization task solved by diffusion of match points. This diffusion is controlled by a potential determined by jet similarity and the topology-preserving constraint. Our algorithm is a simple implementation of the dynamical link architecture for neural networks [1].

1. Beyond Associative Memory:
Object Recognition in the Dynamic Link Architecture

The ability to discriminate and classify objects has to be robust against variations in sensory data. Variations are introduced partly by the nature of the imaging process and partly by variability of objects themselves. The most prominent image variations to deal with are pattern translation, pattern distortion, perspective distortion, variation in background, partial occlusion and changes in pattern size and orientation. More complications arise from changes in lighting conditions, or, in the case of infrared data, of object radiance.

Technical object recognition systems usually avoid all these difficulties by narrowly restricting their domain of application. The effort needed to increase their scope turned out to be enormous. It is therefore necessary to find a generic system architecture which deals with the fundamental problems in a principled way. The nervous system seems to offer proof of existence for such an architecture, and our knowledge about it should be used in the search for a solution to the object recognition problem.

We depart in our discussion from artificial neural systems that are based on the principle of associative storage and recall. Pattern recognition in this context may be based on the following steps:

• take a set of neurons, with linear or nonlinear response behaviour
• design neural interactions such that equilibrium states of the network represent meaningful stored patterns
• prepare your network in an initial state which corresponds to an incomplete or distorted pattern
• let the network relax to the nearest equilibrium state and interpret this state as the pattern recognized.

The performance of systems based on this scheme depends critically on the way patterns are encoded and presented to the system. If neurons just encode local grey levels in the image directly, the network can easily restore pixels distorted by noise, but it completely lacks all invariance with respect to shift, rotation and distortion, which is indispensable for flexible pattern recognition.

Several conceptually different approaches to this problem have been discussed in the literature [2,3,4]. An extensive search through pattern space has been proposed in [3]. One can expect that this approach faces the same problems as template matching algorithms in computer vision — poor convergence due to the size of the search space. An associative memory with a pre-processing unit is discussed and simulated in [4]. Shift- and size-invariance are achieved by consecutive Fourier- and Mellin-transformations. This idea was already proposed in the early seventies [5]. It has the draw-back of being extremely sensitive to image distortions, as produced by perspective transformations, for example. The model proposed in [2] encodes images as graphs and stores them in an associative memory. A preprocessing network realizes all possible permutations of the image graph and matches them to those stored in the associative network. Invariance is obtained by searching through the space of all possible permutations of image points. Thus, the search for good graph matches might be extremely time consuming and computationally expensive. In addition, an enormous number of neurons is necessary to implement the preprocessing (N² neurons for an image composed of N pixels).

The work reported here is based on the Dynamic Link Architecture defined in [1] and discussed in [6,7]. This architecture augments traditional neural nets by the ability to flexibly encode syntactic bindings between semantic atoms represented by single neurons or neuron pools, and to organize complex binding structures efficiently. The basic approach is based on the representation of bindings between semantic atoms by temporal correlations between neural signals. The postulated synchronization is produced by excitatory neural interaction. The process of organization of appropriate binding structures is based on a feed-back loop between signal correlations and rapidly modifying synaptic connections. This feed-back is positive, strong correlation leading to increase in strength of connection, and strong connections causing correlations in neural activity.
This process of self-organization naturally favors certain ordered connectivity patterns which form very useful global data structures. One particular kind of organized connectivity pattern consists of two-dimensional locally connected networks. These are ideal for the representation of objects in an image. An important natural organization process in the Dynamic Link Architecture is the storage and retrieval of network patterns. Another useful process natural to the architecture is labeled graph matching: The ability of the system to "discover" that for an active connectivity pattern there exists another, stored, connectivity pattern which is label-isomorphic (or homeomorphic) to the first, to activate that isomorphic connectivity pattern, and to realize the isomorphism by an activated one-to-one mapping between the two.

The system for object recognition described here is based on labeled graph matching. It makes use of a process which may be conjectured to scale linearly with the number of nodes involved. The system's formulation is close to neural in style.

2. Gabor Transformation of Images

The present system starts with grey-level images, typically in the format \( N^2 = 128 \times 128 \) pixels. The images are sampled, or transformed, with a resolution pyramid of sensitivity functions. In our implementation we chose Gabor functions [8]:

\[
\psi_k(\vec{x}) = \exp(ik \cdot \vec{x}) \exp \left( -\frac{1}{2} \frac{k \cdot \vec{x} \cdot \vec{x}}{\sigma^2} \right). \tag{1}
\]

These functions can be regarded as receptive fields or sensitivity functions of neurons. The first factor is an oscillating function which contains a sine phase (imaginary part) and a cosine phase (real part). The second factor is a Gaussian window localizing the oscillation. The vector \( \vec{x} \) refers to position in the image, and the wave number \( k \) parameterizes size and orientation of the receptive field. The fixed parameter \( \sigma \) controls the size of the Gaussian window relative to the wavelength. All sampling functions can be derived from a single one of them by scaling and rotation.

Frequency space is sampled by restricting the length of \( k \) to the set

\[
k \equiv |k| = \frac{2\pi}{N} (\sqrt{2})^\kappa \quad \text{with} \quad \kappa \in \{1, \ldots, 12\} \tag{2}
\]

and by employing 8 different orientations of Gabor functions, i.e.,

\[
k = k \left( \cos \phi, \sin \phi \right) \quad \text{with} \quad \phi = \frac{\pi \nu}{8}, \quad \nu \in \{0, \ldots, 7\}. \tag{3}
\]

Parameter \( \kappa \) enumerates the different levels of the resolution hierarchy. Values of \( \kappa \) between 7 and 12 define 6 levels of a frequency hierarchy, spanning 3 octaves in total. The width of the Gaussian window was chosen to be \( \sigma = 2\pi \), comprising approximately two oscillations. Each function serves to extract a sample of the image which is localized both in space and in frequency. The sampling functions centered at one particular point cover the two-dimensional frequency plane completely. Each resolution level \( \kappa \) covers a ring-shaped region in frequency space. This frequency band is divided into 8 sectors, each of which is occupied by a Gabor function of one specific orientation. The rings are nested, each spanning half an octave of the frequency range. The sequence of rings breaks off at low and at high frequencies.

The sampling functions, characterized by their frequency, position and orientation, may be considered as feature detectors. Information on whether a specific feature can be found in an image \( I(\vec{x}) \) at position \( \vec{x}_0 \) is measured by convolving the image with the sampling function \( \psi_k \). The resulting Gabor coefficient is defined as

\[
G_k(\vec{x}_0) = \int d\vec{x} \ I(\vec{x}) \ \psi_k(\vec{x}_0 - \vec{x}). \tag{4}
\]

One may reconstruct parts of the image from the information contained in the Gabor components centered on one point \( \vec{x}_0 \), see Fig. 1. Such reconstructions reproduce the image with high resolution near the sample point \( \vec{x}_0 \) and with decreasing resolution further away from \( \vec{x}_0 \).

Each frequency band is evaluated for a discrete square lattice of sampling points in the image. Lattices differ in grid size by half an octave, or a factor of \( \sqrt{2} \), between levels of the hierarchy. The system of lattices thus defined still leaves open a common factor \( k\Delta x = \sigma \). The present system is based on storage of the whole Gabor pyramid of sample points just defined. All coefficients not explicitly represented in the pyramid are reconstructed by interpolation.

The real (imaginary) part of \( G_k(\vec{x}_0) \) in (4) corresponds to cosine (sine) phase of Gabor functions. For a given \( \vec{k} \), \( G_k(\vec{x}_0) \), taken as a function of \( \vec{x}_0 \), is a bandpass filtered version of the image \( I(\vec{x}) \). It thus typically oscillates with a wavelength \( 2\pi/k \). For many purposes, e.g., the labeled graph matching used here, it is highly desirable to represent local information of an image in a smoothly varying way, that is, to avoid rapid oscillations of local features. Therefore, we introduce the magnitudes of Gabor coefficients as

\[
M_k = |G_k| = \sqrt{G_k \cdot G_k^*}. \tag{5}
\]

Let us designate by \( \vec{J}(\vec{x}_0) \) the set of Gabor magnitudes obtained by sampling the image at point \( \vec{x}_0 \) with sampling functions of all sizes (frequencies) and orientations, i.e., the set of \( M_k \) for all possible \( k \) values. We will refer to \( \vec{J}(\vec{x}_0) \) as "Gabor jet."

3. Description of the Pattern Recognition System

Our system for labeled graph matching and invariant pattern recognition consists of two different parts, called image domain (I) and object domain (O). The image domain contains the jets of the Gabor transformation we just defined in (4) and (5). In the object domain of our system we store copies of parts of the image domain for a number of objects to be recognized.

Pattern recognition in our system consists of a dynamical assignment of preselected points in the image to points in stored objects. This dynamical linking of points is guided by a function \( S \) which determines the similarity between an image jet and an object jet. The linking procedure is performed under the constraint that the matching points found in one of the stored objects have approximately the same topological relationship as the preselected points in the image. Introducing this constraint our pattern recognition task is equivalent to the problem of labeled graph matching.

During pattern recognition the system encodes an image by Gabor functions and defines jets for an appropriate subset of matching points. Gabor jets of these selected points should contain enough information for the subsequent topology-conserving map-
The reconstruction is based on the information of 144 Gabor components $\{k \in \{5, \ldots, 13\}, \nu \in \{0, \ldots, 7\}\}$ taken from the central point $(z_0 = (40, 61))$ of the left eye.

The first two factors of $S$ measure the quality of a match, i.e., $S = 1$ if $\mathbf{j}_i^I \equiv \mathbf{j}_i^O$. The third term in (7) weighs the significance of a match, putting emphasis on perfect matches of long jets (rich in information) against perfect matches of short jets.

The cost function (6) is optimized by a standard Monte Carlo method. After choosing a suitable initial configuration of tentative match points — in our case points in object space just identical to those in image space — the cost function is computed. The actual update step of the Monte Carlo procedure consists of two stages:

(i) shift a randomly chosen match point in $I$ and compute the energy difference $\Delta E$ between the old and the new configuration. Then

(ii) if $\Delta E < 0$ the shift is accepted, and if $\Delta E > 0$ it will be accepted with probability $p(\Delta E) = \exp(-\Delta E/T)$.

The parameter $T$, sometimes called computational temperature, measures the stochasticity implicit in the search process. $T = 0$ corresponds to a gradient method. The Monte Carlo dynamics simulates a diffusion of match points in $I$ in a potential shaped by the similarity function $S$. The topology-conserving term in (6) can be interpreted as an interaction between different match points. Both contributions turned out to vary in a smooth way, as can be deduced from the rapid relaxation of the cost values observed. Simulated annealing might be helpful for very noisy images but was unnecessary in our experiments.

As a representative test of distortion tolerance and discriminatory capacity of the proposed labeled graph matching system we have defined the following task and addressed mainly two questions: Take different pictures of the face of one person from different view directions (up to 50 degrees rotation). Does there exist a smooth dependence of final costs on perspective distortions? Does there exist a significant difference in final costs if the images compared show faces of different persons?

The first image (a) in Fig. 2 is taken as a prototype image from the object space $O$ of several stored images. It is compared with three input pictures (b,c,d) taken from the input space $I$. The white grid in image a shows the graph which characterizes this image. For reasons of visualization we have chosen a quadratic
grid instead of a more general topological graph based on salient points. The distorted grids in the other images denote the final configurations of matching points found after sufficient diffusion. The grid distortions correspond to increasing perspective distortions in the images. The final cost values of a match of face $a$ with rotated versions of that face are shown in Fig. 3. Energy increases as a function of the rotation angle. If face $a$ is matched to a different face ($d$) the best match yields significantly higher energy. Reliable classification of faces has been achieved by our graph matching algorithm in all cases investigated.
4. Discussion

The system described demonstrates distortion invariant object recognition. Speaking in technical terms the system may be described as based on local template matching. The templates, inherent in Gabor jets, are designed to be robust with respect to local distortions. Global structure of objects is represented by the constraint of matching topology between arrays of local templates in image space and in object space.

The system implements a special type of labeled subgraph matching. Nodes of graphs correspond to image points or object points. Labels correspond to topological neighborhood relationships. Labels are provided by local features, Gabor functions in our implementation. The general problem of subgraph matching is known to be NP complete. For purposes of image analysis one deals, however, with a special and very benign type of graphs—topological graphs—and the complexity of the problem is moreover reduced decisively by the existence of labels. Consequently the complexity of the computational problem is small.

The flexibility and simplicity of the approach proposed here is illustrated by another, very similar, implementation of the system [9]. That implementation was completed in a few days, and without any domain-specific adaptation the system was able to successfully recognize vehicles in infrared images.

Our present implementation does not insist on neural style in any detail. Floating point numbers were used to make efficient use of present processor structure. A neural representation of jet components would be very costly in terms of processing, and it is not clear whether the system would not work satisfactorily with one or two bits of precision in the jet components anyway.

Also our similarity function $S$ could easily be implemented in terms of threshold functions and synaptic weights, but we did not attempt to do so. A neural implementation of the graph matching itself has been described in [10], and the imposition of the topology constraint would follow the ideas put forward in [11, 7].

Future work will deal with the integration of low-level vision mechanisms into the system, and with improved and adaptable similarity functions.

Acknowledgement: This work has been partially supported by the German Ministry for Science and Technology (ITR-8800-H1) and by AFOSR (88-0274). JB acknowledges support by a NATO grant (DAAD 300/402/513/9).

REFERENCES


