PERGAMON

Neural Networks 12 (1999) 1021–1036

Neural Networks

www.elsevier.com/locate/neunet

Computation of pattern invariance in brain-like structures

S. Ullman[a,*], S. Soloviev[b]

[a]Department of Applied Mathematics & Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel
[b]Department of Neurobiology, The Weizmann Institute of Science, Rehovot 76100, Israel

Received 19 January 1999; accepted 26 March 1999

Abstract

A fundamental capacity of the perceptual systems and the brain in general is to deal with the novel and the unexpected. In vision, we can effortlessly recognize a familiar object under novel viewing conditions, or recognize a new object as a member of a familiar class, such as a house, a face, or a car. This ability to generalize and deal efficiently with novel stimuli has long been considered a challenging example of brain-like computation that proved extremely difficult to replicate in artificial systems. In this paper we present an approach to generalization and invariant recognition. We focus our discussion on the problem of invariance to position in the visual field, but also sketch how similar principles could apply to other domains.

The approach is based on the use of a large repertoire of partial generalizations that are built upon past experience. In the case of shift invariance, visual patterns are described as the conjunction of multiple overlapping image fragments. The invariance to the more primitive fragments is built into the system by past experience. Shift invariance of complex shapes is obtained from the invariance of their constituent fragments. We study by simulations aspects of this shift invariance method and then consider its extensions to invariant perception and classification by brain-like structures. © 1999 Elsevier Science Ltd. All rights reserved.

Keywords: Shift invariance; Pattern invariance; Object recognition; Visual system

1. The problem of shift invariance

Our visual system can effortlessly recognize familiar objects despite large changes in their retinal images. The image of a given object changes due to variations in the viewing conditions, for example, changes in the viewing direction, illumination, position and distance. The visual system can somehow compensate for these changes and treat different images as representing an unchanging object. Many of the images we see are novel either because they depict objects not seen before, or because familiar objects are seen under a new combination of viewing conditions. Yet, the visual system can use its past experience with the same or similar objects to correctly classify and recognize the viewed objects.

One of the image transformations that the visual system can compensate for is a change in the retinal position of the viewed object. This is a relatively simple transformation, and yet there are no satisfactory and biologically plausible models of how shift invariance is obtained in the brain. In this section we review briefly the main approaches to shift invariance and their shortcomings, together with the main relevant psychophysical and physiological findings.

1.1. Main approaches to the modeling of shift invariance

An image that falls at different retinal locations is initially registered and analyzed, at the levels of the retina, LGN, and primary visual cortex, by different sets of neuronal mechanisms. To achieve shift-invariant perception, these different initial representations presumably reach a common unified representation at some higher levels of the visual system. How is this obtained by the circuitry of the brain? Is this an innate or an acquired capacity? If it has an acquired component, must it be learned for each object individually, or can it be generalized from some objects to others?

Perhaps the simplest (but highly redundant) approach to obtaining shift invariance is by what may be called "full replication". According to this approach, a specialized neuronal mechanism is dedicated to the detection of a given shape, such as the letter "A" of a particular shape, at a given position in the visual field. To recognize the same shape at different locations, multiple replications of the same detection mechanism are used. The detectors at the different locations can then converge upon higher order

* Corresponding author. Tel.: + 972-8-934-2894; fax: + 972-8-934-2945.

0893-6080/99/$ - see front matter © 1999 Elsevier Science Ltd. All rights reserved.
PII: S0893-6080(99)00048-9

cortical units, which will exhibit shape-specific, but position-invariant response.

Some approaches used replication of shape-detecting mechanisms but at different levels of complexity (Földiák, 1991; Fukushima, 1980). Invariance is achieved by first detecting simple image features, such as line segments, and then combining them to create increasingly complex feature detectors. For example, a *T*-shape can be detected by combining horizontal and vertical line detectors within a small neighborhood. A problem with this scheme is that the spatial relation between the features is lost in the combination. In the scheme we develop in Section 2 shift invariance for complex shapes is also based on the shift invariance of simpler components but in a manner that avoids this and a number of other shortcomings.

At the opposite extreme to this highly redundant approach is the normalized representation approach. The general idea behind this approach is to transform the input image into a normalized central representation, common to all retinal positions, and let the pattern analyzing mechanisms operate upon this common representation. This general approach is frequently used in artificial computer vision systems. A biological model using a common normalized representation was proposed by Olshausen, Anderson, and Van Essen (1993, 1995) based on attention-dependent mechanisms. The model is able to handle position and scale invariance by shifting and scaling image regions between input and output arrays. A region in the image is selected by an attentional window, and the sub-image inside this window is mapped onto an object-centered reference frame, regardless of its shape and size. This is obtained by a "dynamic routing" circuit that controls the connection strengths between input and output layers. A shortcoming of this model is that the scheme requires a complex network with unrealistic switching mechanism of individual synapses. A second shortcoming is that the model does not generalize naturally to other invariances such as rotations in space. A third problem of the normalized representation approach in general is that it implies shift invariance for arbitrary novel shapes. As we shall see in the following section, this property is inconsistent with psychophysical results (Dill & Fahle, 1997; Nazir & O'Regan, 1990) that reveal significant limitations of shift invariance for novel, complex shapes.

Other models can often be viewed as lying at intermediate points between full replication and a single normalized representation. For example, back-propagation networks were trained to obtain shift invariance in particular application domains such as character recognition (Le Cun et al., 1989). The hidden layers is these networks develop intermediate representations that are more compact than full replication, but do not produce a single normalized representation of the input. It turned out that shift invariance is difficult to obtain in such networks. Special mechanisms such as weight-sharing that is biologically implausible were sometimes incorporated to endow the networks with a higher degree of shift invariance. In conclusion, none of the mechanisms proposed so far can obtain shift invariance in a computationally efficient and biologically plausible manner.

1.2. Empirical evidence

In this section we review briefly the main empirical evidence, both psychophysical and physiological, related to models of shift invariance. We also include a brief comparison with evidence regarding size invariance in object recognition. The basic characteristics of these findings will be used in the model described in the subsequent section.

1.2.1. Psychophysical studies

Psychophysical studies have shown both the power and the limitations of shift invariance in the human visual system. A number of studies have shown that under favorable conditions considerable position invariance can be obtained. For example, Biederman and Cooper (1991) obtained evidence for a high degree of position invariance for line drawings of familiar objects. The evidence was based on priming experiments, showing that the amount of priming was unaffected by translation of five degrees in the visual field. Bricolo and Bülthoff (1992) also obtained evidence that humans can recognize images of shaded three-dimensional objects at new retinal locations.

With respect to changes in size, priming experiments by Fiser and Biederman (1995) also showed that the decrease in reaction time following priming was independent of whether the primed picture of a familiar object was presented at the same size as the original picture or at a different size. However, when testing the effects of size on recognition using a same/different task, a systematic size dependence for reaction times and error rates was observed. A similar increase of recognition latencies with the discrepancy in size between learned and viewed shapes was also found by Bricolo and Bülthoff (1993), and by Jolicoeur (1987).

Experiments that measured shift invariance more directly, especially using complex and novel stimuli, revealed substantial limitations on the degree of shift invariance exhibited by the visual system. For example, in experiments by Nazir and O'Regan (1990), subjects were trained on the discrimination of novel patterns at one location, and then tested at a nearby (0.49–2.4°) location. Significant decrease in performance was found at the new location. Similar results were obtained recently by Dill and Fahle (1997) and Dill and Edelman (1997). Using somewhat different methodology and patterns, they also found that extensive training to discriminate between similar novel patterns at one location did not improve performance when tested at a new location. It was also found that for isolated figures the location to which the subject attended did not affect the performance in their shift-invariance tests, which is significant for some theories that rely on directing attention to achieve shift-invariant recognition (Olshausen et al., 1993; Salinas & Abbot, 1997).

In conclusion, it appears that considerable shift invariance can be obtained with highly familiar shapes and when fine discrimination between similar shapes is not required. For novel shapes, when fine discrimination is required, shift invariance in considerably reduced. The experiments suggest that shift invariance is not automatic and universal. It requires training with specific patterns that must be presented at multiple locations. It also appears that there is an effect of generalization across patterns, namely, training for a set of patterns can improve the performance for similar patterns. In the experiments of Dill and Fahle, for instance, the performance in discriminating the novel patterns was above chance from the outset, presumably because of past training with other patterns. Similarly, novel images that are new variations of highly familiar objects (as in the experiments by Biederman and Cooper) tend to produce a relatively high degree of shift invariance.

1.2.2. Shift and size invariance in the visual system

From physiological studies of receptive field properties along the primate visual pathway it appears that shift and size invariance is built gradually from the primary visual cortex to high-level visual areas in the infero-temporal cortex. A limited degree of shift invariance for oriented line and edge stimuli is observed already at the level of the primary visual cortex, area V1. As was shown by Hubel and Weisel (1962), complex cells in V1 exhibit shift-invariant responses to their preferred stimuli over a limited range.

At higher stations along the visual pathway, units typically exhibit specificity to more complex stimuli, and this specificity is often maintained over increasingly large receptive fields. Units in area V4 were shown to respond optimally to more complex local patterns than simple linear features, including radial, concentric, and hyperbolic grating patches (Gallant, Braun & Essen, 1993). Most of these units show similar response magnitude and specificity across their receptive fields. Lesion studies in V4 by Schiller and Lee (1991) suggest that this cortical area plays an important role in obtaining translation invariance of complex shapes, since a lesion to this area has significant effects on the animal's ability to recognize a familiar shape at new locations. Similar results were obtained in a subsequent study (Shiller, 1995) with respect to size invariant recognition.

The tendency to create units with more complex shape specificity and larger position invariance continues in the areas of the infero-temporal cortex (Desimone, Albright, Gross & Bruce, 1984; Gross, 1992; Gross & Mishkin, 1977; Ito, Tamura, Fujita & Tanaka, 1995; Perrett, Rolls & Caan, 1982; Tanaka, 1993; Tanaka, Saito, Fukada & Moriya, 1991). For example, Desimone et al. (1984) found that stimulus selectivity of IT cells was maintained over substantial changes in stimulus position as well as size; cells responded to face images ranging in size from 2.5 to 10°, anywhere within receptive fields as large as 40–60° centered on the fovea. Tanaka and coworkers (1991, 1993) found columns in IT that contain cells responsive to complex visual features. The receptive fields of these cells were large (average square root of the area 13.62 ± 7.32°), and many showed comparable responses across their receptive fields, and changes in stimulus size as large as eightfold. Ito et al. (1995) showed that anterior IT cells exhibit significant invariance to changes in the retinal images of objects in position as well as size. A minority of the cells in IT are more position and size specific, for example, the response decreases significantly with doubling the stimulus size. Using a different experimental method, fMRI brain imaging of human subjects, Malach, Grill-Spector, Edelman, Itzchak and Kushnir (1998) obtained evidence that a cortical region termed the lateral occipital complex (LO), that has been implicated in object perception, exhibits considerable shift and size invariance for grey-level pictures of complex objects.

The results suggest that shift invariance emerges in the hierarchy of visual areas in the primate visual system in a series of successive stages, and it is accompanied by an increase in the complexity of the shapes analyzed by these units. Units in the primary visual cortex respond to simple visual features such as lines and edges over a limited region. Units in areas of the infero-temporal cortex respond to complex patterns and views of object-parts and complete objects over large portions of the visual field. It is unclear, however, how this elaboration of shape specificity together with increased position tolerance is obtained in the stream of processing from low to higher visual areas. In the following sections we propose and test a model that naturally incorporates these aspects of shape analysis by the visual system.

2. Shift invariance by the conjunction of fragments

We have seen above the limitations of both the full replication and the single representation approaches to the problem of shift-invariant recognition. The full replication model is straightforward, and it uses the brain's inherent parallelism and the existence of multiple units responding selectively to a variety of different shapes. At the same time, the proposal to have a separate mechanism at each location tuned to each recognizable image is implausible because of its extreme redundancy and the limited ability to generalize to new patterns. As for the single representation approach, the normalization process appears biologically implausible, and this approach does not account for the main properties of units along the visual pathway, and does not account for the role of learning in obtaining shift invariance for novel families of stimuli.

Our proposed approach to shift-invariant recognition (as well as to other aspects of invariant recognition) is an

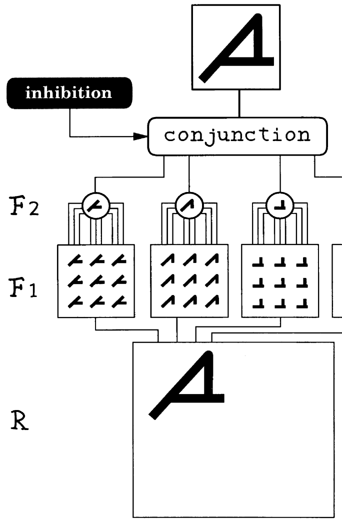

Fig. 1. Detecting a shape by the conjunction of overlapping fragments. F2 unit responds to the maximal activity of its F1 units.

intermediate one. It uses full replication, but at the level of object fragments rather than complete views. These shape fragments can then serve as building blocks for defining much larger sets of complex shapes. According to this view, the brain will learn over time to extract appropriate shape fragments, as well as the connection between similar fragments at different locations. Shift invariance for complex shapes is then obtained from straightforward conjunctions of the responses to the more elementary fragments comprising the full shape.

The fragment detectors are simpler than full-view detectors, but more complex than elementary feature detectors such as an oriented edge. In this scheme, not all object representations are stored at each position, but only a number of partial fragments, at a number of complexity levels. As we shall see, view-fragments of this type can be used in such a manner that a relatively small number of features can allow the invariant recognition of a much larger set of complex patterns.

A generalization of this approach is also proposed for other forms of invariant recognition. The general proposal is that the brain constructs, on the basis of past experience, a

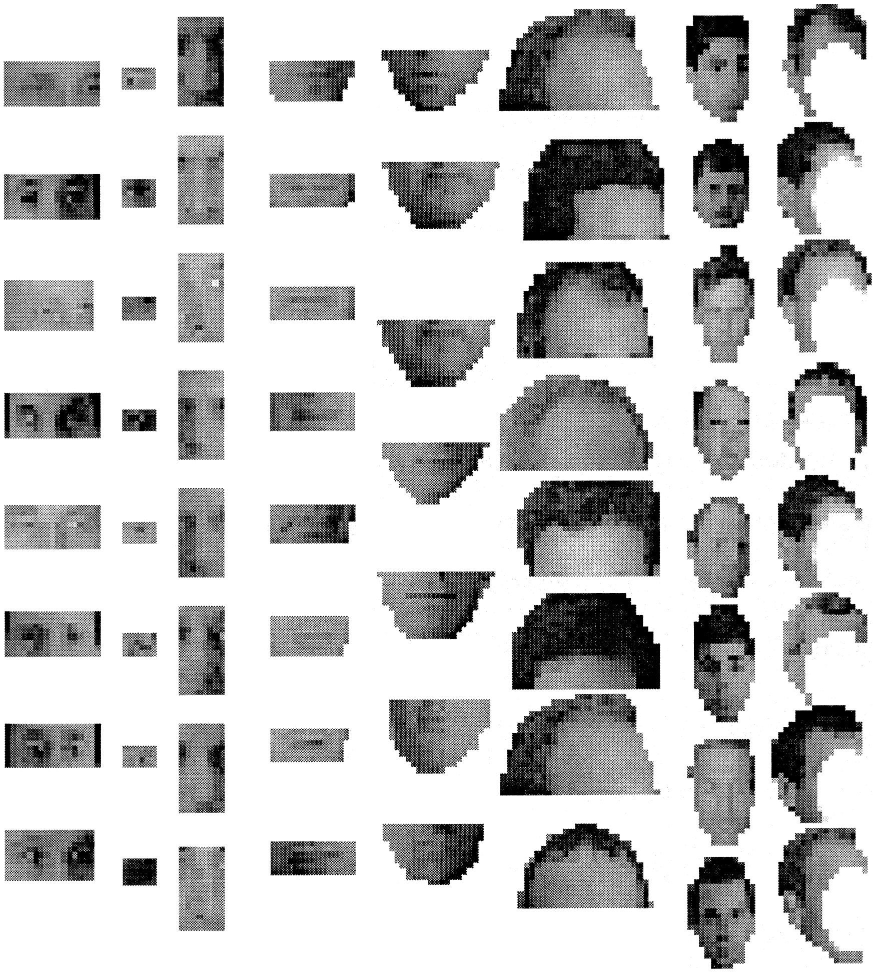

Fig. 2. Example of face parts from a system for face detection using the conjunction of fragments. The system uses multiple overlapping fragments, at different levels of resolution. Figure prepared by E. Sali.

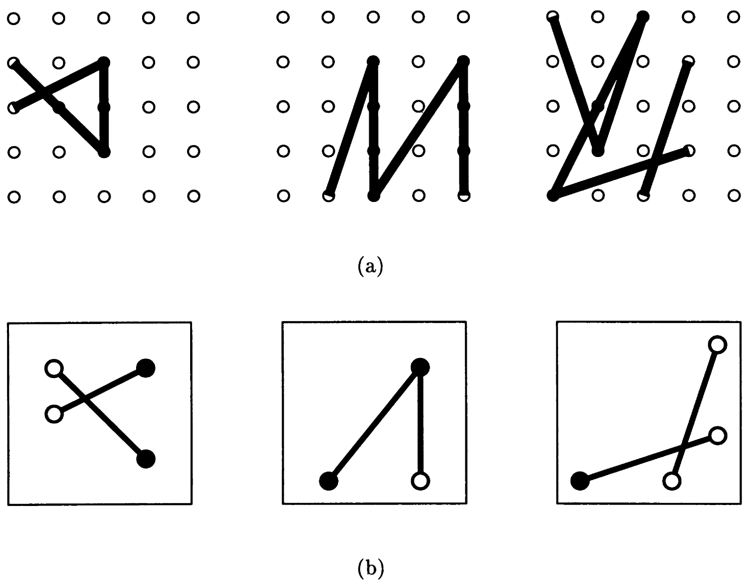

Fig. 3. Shapes and shape fragments: (a) examples of input shapes; (b) examples of fragments, ○ denotes free ending ● denotes a junction.

large repertoire of partial generalizations. It stores view-fragments of different complexity as well as the equivalence relations among them. This stored repertoire is then used in processes involved in invariant recognition, generalization, and object classification.

A schematic illustration of the conjunction of overlapping fragments is shown in Fig. 1. An input shape is to be detected anywhere within a neighborhood (R in the figure). Different partial fragments of the shape are detected by units tuned to different partial shapes, and replicated at different locations (F_1 in the figure). The first stage towards position invariance is obtained by the convergence of these units to generate fragment detecting units that are position invariant within their receptive fields (F_2 in the figure).

The detection of the full shape is then obtained by the conjoint activation of the constituent fragments. Inhibition can also be used to preclude the existence within the receptive field of fragments that do not belong to the input shape. As we shall see, it is crucial for the proper operation of this scheme that the shape will be covered by a redundant and overlapping set of fragments. These properties enforce the correct spatial arrangement of the constituent fragment. Another aspect of this scheme is that it is not possible in general to create the invariant fragments in a single step. Fragments are constructed in a number of steps, and at each successive step units respond to increasingly complex shapes and over larger regions in the visual field.

The most fundamental problem that arises in attempting to derive invariance to complex shapes from the invariance to simpler components, is the problem of spatial relations. It may appear that the mere conjunction of the constituent fragments will be insufficient for recognizing correctly the full shape. This is because the spatial relations among the fragments are not specified explicitly. This may cause the recognition system to confuse two different shapes that are comprised of identical fragments, but in a different spatial arrangement. Spatial relations between components can be captured, however, by the use of multiple, overlapping fragments. For example, in representing an image of a face by the conjunction of fragments, the fragments will not be limited to the use of a small number of natural, disjoint parts such as the eyes, nose, mouth, etc., as in other proposed schemes (Biederman, 1985). Instead, multiple overlapping fragments will be used, including fragments covering, for instance, a part of an eye with a portion of the nose, internal features with portions of the bounding contour, etc. Fig. 2 illustrates examples of face-fragments from a model under development for face detection based on the conjunction of fragments. If the face figure is rearranged, so that the face parts appear in a jumbled configuration, some of the parts such as the eye or nose may still be detected, but many of the other fragments will no longer be present.

The use of multiple overlapping fragments to implicitly code for spatial relations, and overcome the problem of part rearrangement, is illustrated in the next section. This section describes the application of the scheme for shift-invariant recognition in a domain of simplified line drawings. As we shall see, in this domain a small number of simple fragments can be used for uniquely defining, in a shift-invariant manner, a much larger set of more complex shapes.

2.1. Shift-invariant recognition of simple line shapes

As discussed in the preceding section, we suggest that shift invariance can emerge in a highly parallel network of localized detectors for a sufficiently rich family of overlapping shape fragments. Low-level layers of this network contain at each location a set of detectors tuned to different partial shapes. These partial shapes, or fragments, are not

Table 1
Non-uniqueness results for 3 × 3 input shapes and 736 shift-invariant fragment detectors

Input shapes of size 3 × 3	
Number of shapes (n)	10 003
Number of fragments	736
Non-unique shapes	16
Standard deviation	0.0004
Fraction of non-uniqueness ($\bar{x} \pm 1.96\sigma_{\bar{x}}$)	0.0016 ± 0.0008

universal in nature, but depend on the set of images to be recognized. For example, the set of shape fragments used to recognize written characters will be different from the fragments used to recognize face images. At subsequent layers in the network the localized fragment detectors are combined to create position-invariant detecting units. Finally, complex shapes are detected by the conjoint activation of the fragment detecting units.

We tested the scheme of achieving shift invariance by the conjunction of shape fragments in two domains — simple line shapes, and grey-level patches. The main issues addressed by these simulations are the following. First, that shift-invariant recognition can be obtained by the co-activation of the constituent fragments, without explicitly coding for their spatial relations. Second, the issue of generalization, namely, that shift invariance can be obtained for novel shapes, provided that they share some similarity with known shapes. Finally, the simulations examined aspects of efficiency, for example, how large is the set of shape-fragments needed to encode a large set of input shapes.

In the domain of line shapes, the input shapes we considered were line drawings generated by connecting pairs of points on a square grid, somewhat analogous to letter-like figures and other simple line drawings. Examples of such shapes are shown in Fig. 3(a). The shapes were restricted to be connected figures, rather than a random collection of lines, and we tested shape families of different sizes, e.g. shapes that are up to 3 × 3 or 4 × 4 grid elements. The task of the simulated network is to recognize correctly the input shapes falling anywhere within a much larger input grid. We assume at each point the existence of a set of simple part detectors. The parts were limited to connected pairs of line segments, as shown in Fig. 3(b). These simple parts are also assumed to contain information about line ending, analogous to "end inhibition" in primary cortex units. That is, line endings within a part are divided into either a free termination or a part of a more extended line. The number

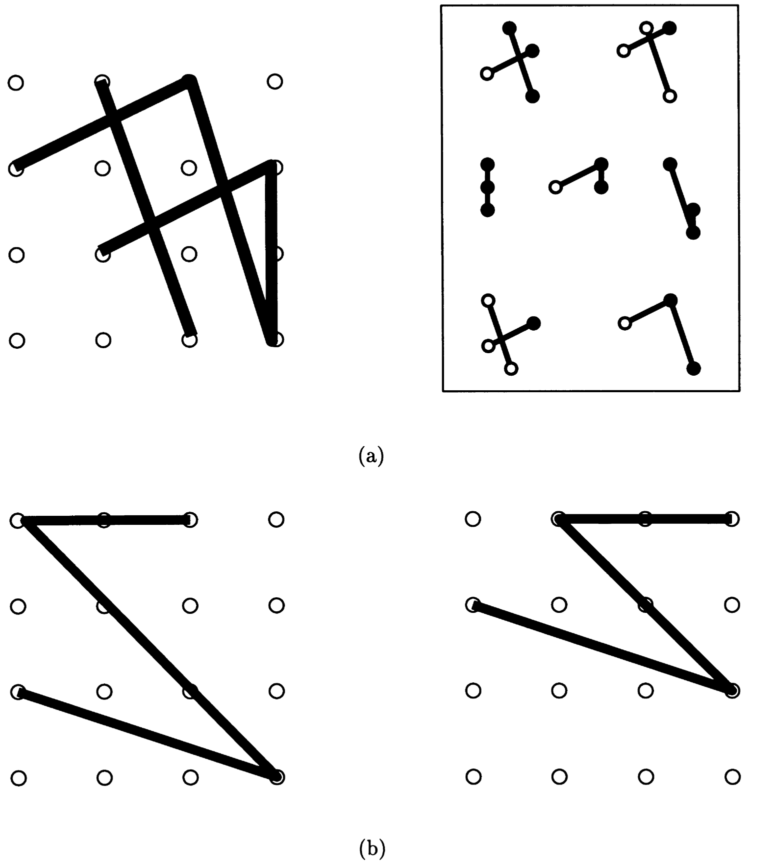

Fig. 4. Unique and non-unique decomposition: (a) a collection of fragments and the unique shape composed of these fragments, ○ denotes free ending, ● denotes a junction; (b) two different shapes with the same fragment composition.

Table 2
Results for a 4 × 4 input grid

Input shapes on a 4 × 4 grid	
Number of shapes	10 596
Number of non-unique shapes	0

of parts we used was small compared with the number of shapes tested. Mathematically, the number of different shapes of size $n \times n$ grows exponentially with n^2. Interestingly, it can be shown that the same holds true for the set of connected rather than all line figures. In contrast, the number of possible parts of the type we used grows only polynomially in n. In the case of 3 × 3 shapes we studied, we used several hundred parts to deal with tens of thousands of input shapes.

To generate parts, we first generated a set of line figures, and then simply used all the parts, that is, connected line-pairs that appeared in these figures, as well as rotated versions of these parts. We next generated a much larger set of line shapes, and tested whether these shapes were defined uniquely by the combination of their fragments, without any additional information regarding spatial arrangement. To this end, we developed a search algorithm that can take as input a collection of parts, and produce all the possible (connected) line shapes that are composed of this set of parts. In this manner we could take an input line-shape, and first find its constituent parts. The parts typically covered the input shapes in a redundant, overlapping manner, namely, each segment in the shape was covered by more than a single part. We then tested whether the same collection of parts could be used to generate a different shape that has the same local fragments but arranged globally in a different configuration.

Statistical results for a sample of about 10 000 input shapes of size 3 × 3 are shown in Table 1 (from Soloviev, 1997). The shift-invariant system was tested on a sample of different randomly generated input shapes to find the degree of its non-uniqueness, that is, the fraction of different inputs having the same representation in the space defined by the shift-invariant fragments. The fragments were defined as pairs of input edges having two types of joints—”free” and “non-free”. Overall, a total of 736 fragments were used. The results illustrate that unique decomposition is almost always guaranteed. Using a set of several hundred shape fragments, a much larger set of inputs can be defined uniquely by the conjunction of their constituent fragments. Only in a small fraction of the shapes, some ambiguity remains. The last two rows estimate the deviations one can expect from the experimental result. (The standard deviation was determined under the simplified assumption of independent Bernoulli trials, and the last row is an estimation of the 95% interval.) An example of unique and non-unique shapes is shown in Fig. 4. It can be seen that the two shapes having the same fragment structure are visually similar, this was typical of the non-unique cases we have examined. Uniqueness can be further enhanced, if desired, by extending the set of fragments. This can be done in two ways. First, specific new fragments can be added to the system to deal with specific ambiguities. This raises interesting questions for further studies, both computational and psychological, regarding incremental learning procedures by the system, to improve its invariant recognition. Second, as can be expected, we also verified that by using from the outset a more discriminative set of fragments, the fraction of non-uniquely defined shapes decreases. Table 2 shows an example using fragments defined with four different types of joints—“free end”, “corner”, “straight”, and “junction”. The size of the shapes was larger in this test (4 × 4), posing a more challenging task, but of over 10 000 input shapes tested, all were uniquely defined in terms of the conjunction of their constituent fragments.

We conclude that the joint activation of the fragments composing a shape is powerful by itself, without additional explicit information regarding relative spatial relationships, to encode large sets of shapes with high degree of specificity. Further, a limited set of fragments can encode uniquely a large set of more complex shapes. If the system contains shift-invariant units for the basic fragments, it will also exhibit shift invariance for novel complex shapes, as long as the new shapes are constructed from existing shift-invariant “building blocks”. For completely novel family of shapes that are not expressed well by existing fragments, shift invariance will be limited. It can improve with practice as the system learns to extract and use additional fragments, derived from the new family of shapes.

The example outlined above used a single step, from fragments to complete figures. When the figures are larger, and when the connectivity assumption used in the line shapes is not used, it becomes natural to construct the fragments in a number of successive steps, where at each step the intermediate shapes become larger and more complex. A simple example of this multi-stage construction is illustrated next, in the domain of binary patches taken from real images.

2.2. Shift-invariant recognition of image patches

In the human brain, shift-invariant units appear to be constructed hierarchically in a number of successive stages, where at each stage the units generalize over larger regions of space, and respond to more complex configurations. We wished therefore to examine whether the hierarchical construction of increasingly larger units offers an advantage over possible alternatives in which large units are constructed from smaller ones in a single step. In this part, we used small image patches obtained from real grey-level images of different objects. By using these patterns we also wanted to get some information regarding the nature of small image fragments that appear in natural images.

To simplify the analysis we first transformed the grey-level images into binary ones, using the following

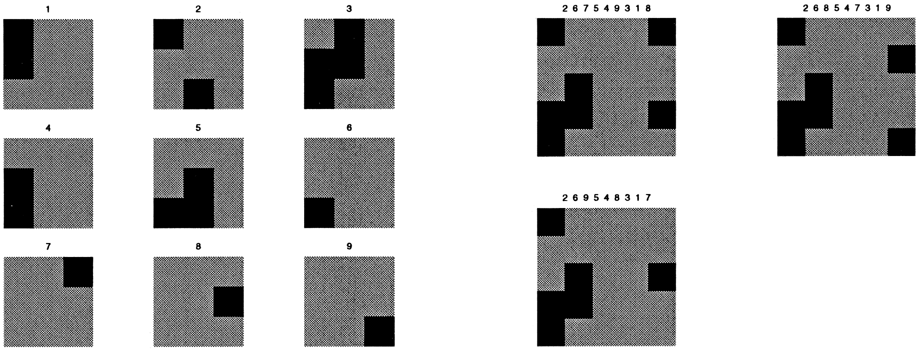

Fig. 5. Nine 3 × 3 micro-patterns and three different 5 × 5 patterns composed of these micro-patterns.

procedure. The input images were convolved with an edge detecting filter shaped as the Laplassian of a Gaussian,

$$H = (\nabla^2 G)I.$$

This is a standard filter used in image processing, and it is also an approximation of the receptive field shapes of retinal ganglion cells (Marr & Hildreth, 1980). Binary images are then obtained by using the sign of the filtered image, that is, the negative values are all set to black, and the positive values to white. This simplification also has the advantage that the resulting image no longer depends on absolute grey levels and is less dependent upon small variations in the illumination conditions.

We started the construction of fragments by using small micro-patterns of 3 × 3 patches. The number of such micro-patterns is limited (512), so we assumed that all or most of these patterns can be detected and used for the construction of larger units. The small fragments are used next to construct larger micro-patterns, of size 5 × 5. A single 5 × 5 patch contains within it nine smaller 3 × 3 patterns. One question that arises in light of the discussion so far is whether the larger patterns in this case are unambiguously determined in terms of the joint activation of the constituent sub-units. Mathematically, the question is whether nine micro-patterns taken from a 5 × 5 patch can be put together in more than a single arrangement. To answer this question, a depth first search (DFS) algorithm was applied to the tree of possible arrangements. The search is heavily constrained, because fragments must overlap properly. After execution of the DFS, the number of possible different reconstructions was recorded. Using a sample of about 17 600 patterns taken from real images, we found that about 3.4% of the 5 × 5 patterns proved to have a non-unique representation in terms of their nine 3 × 3 micro-patterns. This reconstruction is performed for analyzing the feasibility of constructing invariant units in this manner, it does not imply, of course, that the brain must use a reconstruction stage in order to reach invariant recognition. We conclude that the representation in terms of the small micro-patterns is unique for the large majority of the patterns. An example of a non-unique representation is shown in Fig. 5. The figure shows three different 5 × 5 patterns that have the same composition in terms of their smaller fragments. If desired, this ambiguity can be further reduced by incorporating a small number of additional fragments. It also turned out that in this case ambiguity can be avoided entirely by using somewhat different decomposition: if elongated 5 × 2 micro-patterns are used (four horizontal, four vertical), complete uniqueness is obtained. As summarized in Table 3, in most cases the representation of 5 × 5 patterns taken from natural objects by micro-patterns is unique.

Table 3
Results on non-uniqueness (all possible reconstructions by the DFS)

Percent on non-uniquely represented 5 × 5 patterns	
Micro-patterns	Non-uniqueness (%)
Nine 3 × 3	3.4 ± 0.3
Eight 5 × 2/2 × 5	0
Nine 3 × 3 + four 5 × 2/2 × 5	0.9 ± 0.3

Because the larger patterns (5 × 5 in our schematic example) are determined uniquely by the conjunction of their constituent micro-patterns, it becomes possible to construct shift-invariant units for the larger patterns by a convergence of the more elementary sub-units within a region. This can lead, as discussed further below, to the emergence of intermediate units, that will respond, in our schematic example, to a given micro-pattern over a 5 × 5 region.

The next issue we wish to examine is to compare the construction of larger image fragments either hierarchically or non-hierarchically. In the non-hierarchical method, large units are composed directly from the small micro-patterns, and in a hierarchical scheme intermediate units are used. For the current argument, the crucial test is whether the larger patterns can be specified by the combined activation of their participating micro-patterns. As it turns out, this representation becomes increasingly ambiguous as the size of the patterns is increased. For example, for 6 × 6 patterns (that contain 16 micro-patterns) the level of ambiguity is 5%, and for 7 × 7 patterns (containing 25 micro-patterns) the ambiguity increases to about 13% of the patterns. These results were obtained using the DFS algorithm that considers all possible patterns. It is also of interest to consider not all

Table 4
Ambiguity for the non-hierarchical representation

Percent of non-unique fragments in several 3×3		
Method	Result (%)	
	With R	Without R
5×5 in 9 3×3	0.39 ± 0.09	0.39 ± 0.09
6×6 in 16 3×3	0.48 ± 0.09	2.07 ± 0.19
10×10 in 64 3×3	1.19 ± 0.13	6.99 ± 0.30

possible patterns, but only "natural" patterns that arise from natural images. We used a sample of about 20 000 patterns obtained from natural images, and tested them for uniqueness. A pattern was considered ambiguous if it shared the same representation in terms of micro-patterns with a different pattern in our sample.

The results are shown in Table 4. The table also compares two possible variations of specifying larger units in terms of smaller building blocks. A small micro-pattern may occur within the large pattern more than once. The system may represent the number of occurrences for each of the micro-patterns, or, in a simpler representation, just the presence or absence of a micro-pattern. The first possibility is denoted in the table as "with R" (with repetitions), and the simpler representation as "without R".

The results concerning ambiguity are compared with the hierarchical construction of large (10×10) patterns from micro-patterns via intermediate (5×5) patterns. As shown in Table 5, in the hierarchical construction the ambiguity is significantly reduced. If all the participating fragments are used, the ambiguity is eliminated. It is also possible to use a fraction of the fragments (only nine out of the 36 sub-patterns), without compromising much in terms of ambiguity. As before, it is also interesting to note that the remaining ambiguous patterns are visually quite similar. It is of interest to note in this regard that the construction from sub-units offers a natural similarity measure between patterns, which is their degree of overlap in terms of the constituent fragments.

The results indicate that the convergence, in a single step, of small image-fragments to define considerably larger fragments can result in ambiguous units that respond in a similar manner to several different patterns. The construction of invariant units becomes more robust if it is performed hierarchically: units responding to small fragments converge to create intermediate units, which in turn converge to create larger units. The system will eventually contain units encoding fragments of different size and complexity.

Table 5
Ambiguity for the hierarchical representation. Uniqueness was tested on a sample of 7600 patterns, and by a comprehensive depth-first search (DFS, last row)

Percent of non-uniquely represented 10×10	
Method	Results (%)
36 5×5 (on a sample)	0
9 5×5 (on a sample)	0.09 ± 0.07
9 5×5 (DFS)	0.6 ± 0.2

We have seen in the discussion above how we can use the local convergence of different patterns within a restricted region to create well-defined larger fragments. For the purpose of shift-invariant recognition it is also useful, however, to bring together the responses of a set of identical fragments from a much larger region of the visual field.

As a simplified concrete example, consider the detection of a particular target pattern, say of size 5×5, at different retinal locations within a large region, by the combined activity of smaller 3×3 micro-patterns. This can be accomplished in a straightforward manner by assuming the existence of detectors for a particular type of micro-pattern, but at different locations, that is created by the convergence of the micro-patterns onto a single detection unit. This unit, which we will refer to as a shift-invariant fragment, will respond to the presence of the micro-pattern in question anywhere within the large region. The creation of such units depends on a prior learning stage by the system. Biologically plausible models have been described for the creation of such units, that emerge by simple synaptic modification rules as a result of patterns drifting across the retina (Földiák, 1991; Parga & Rolls, 1998). In this manner the local fragment-detectors converge onto a more global unit that responds to the presence of a particular fragment anywhere within a large region of the visual field. As a result, the presence of the target pattern anywhere within the region will activate all the micro-patterns it contains.

In this scheme, a novel target pattern, seen only at one location, can be recognized at other locations as well. The target will activate at the two locations a similar set of micro-patterns. If the micro-patterns at the two locations converge, as proposed, onto the same shift-invariant higher level fragment detectors, the system will be able to immediately generalize from the initial to the new location.

It appears natural to combine such a scheme of converging sub-units with a mechanism that limits the region in the visual field within which information is collected for the purpose of recognition, especially when the pattern to be recognized is embedded within a larger pattern. Within the context of the larger pattern, many additional micro-pattern detectors will be activated. This complicates the task and may lead to "false conjunctions" (Treisman & Gelade, 1980) if all or most of the micro-patterns are present but not in the correct configuration. If the analysis of the pattern is restricted, however, to a region around the target pattern, the simulations indicate that possible misidentification will be reduced, or eliminated entirely if the analyzed region is sufficiently small.

Reliable recognition at the new location, especially for embedded figures, can be aided therefore by segmentation or attentional processes that restrict the gathering of relevant information to the neighborhood of the pattern to be analyzed. If similar principles operate in the brain, it is not surprising that in human vision recognition also depends on image segmentation and attentional processes (Ullman, 1996).

Let us summarize our conclusions regarding the hierarchical construction of units, first in the framework of the small grey-level patches we have considered as an example, and then in a more general manner. We have seen in the discussion above that both local and global convergence of the fragments play a useful role in obtaining shift-invariant recognition. Regarding global convergence, we have seen how the shift-invariant recognition of a pattern can be obtained within a large region using the combined activity of the smaller fragments it contains. For example, a 5×5 patch can be detected in a shift-invariant manner by the combined activity of the 3×3 micro-patterns. For this purpose it is convenient to use global convergence of the micro-patterns combined with a segmentation or attentional process that restricts the region over which information is gathered.

We have also seen that for the recognition of considerably larger patterns (10×10 in our example), it is not enough to use information from the elementary micro-patterns, but fragments of intermediate size must be used. Such intermediate units can be created by the local convergence of smaller micro-patterns. It then becomes possible to recognize large patterns in a shift-invariant manner as well. In our example, 5×5 intermediate units are used to recognize reliably the large patterns, and these intermediate units by themselves are constructed from the local convergence, within their receptive fields, of the more primitive micro-patterns. The elementary micro-patterns are therefore used in the system to create higher-order units in two complementary ways. First, micro-patterns of different types are combined locally to define more complex local fragments. Second, micro-patterns of the same type are combined more globally to create shift-invariant units that respond to the micro-pattern over a large region. This is an inherent aspect of the system's architecture, and we can therefore expect similar types of unit to exist in the human visual system as well.

We conclude that the construction of invariant units becomes more robust if it is performed hierarchically: units responding to small fragments converge to create intermediate units, which in turn converge to create larger units. The system will eventually contain units encoding fragments of different size and complexity. This hierarchy of fragments can be used to recognize in a shift-invariant manner patterns of any desired size and complexity. The main challenge confronted by a shift-invariant recognition system is to recognize correctly a novel shape, seen at one location, when presented again at a new location. Within the fragment-based scheme, the shift-invariant recognition of a novel shape will be obtained by the combined activity of the partial fragments activated by the shape. Full invariance will be obtained for novel shapes, provided that the shape in question can be encoded in terms of the existing fragments. If the encoding in terms of existing fragments is only partial, and does not encode the novel shape uniquely, then the invariance exhibited by the system will not be complete: for example, it may confuse a given shape with a similar shape at a different location.

3. Computation of pattern invariance in brain-like structures

In this section we first summarize the main properties of the approach to shift invariance and its implications, and then discuss the application of a similar approach to other aspects of invariant pattern perception.

3.1. Shift invariance

The main goal of the approach discussed so far is not to develop a detailed biological model for shift invariance, but to outline an approach for dealing with more general aspects of invariant recognition and generalization in brain-like structures. We will therefore summarize only some basic aspects of shift invariance and then discuss their implications and possible extensions.

The basic model is that shift invariance of shapes is obtained by the conjunction of more primitive shift-invariant building blocks. Unlike the cortical shift theory, this approach does not rely on a single canonical representation, but uses multiple replications of similar units at different locations in the visual field. However, replication takes place not at the level of the complete patterns that are being recognized, but at the level of primitive image fragments. The construction of fragments is performed hierarchically, in a number of stages. The fragments used are specific to a family of shapes, unlike previous schemes that used universal simple shape elements such as line segments or Gabor patches.

In any scheme that uses fragments, a basic issue that arises is the problem of rearrangement. That is, the system should not confuse a given shape with a different one that is constructed from the same fragments, but arranged in a different configuration. The basic mechanism used in the proposed scheme to impose the correct spatial arrangement is the employment of a rich enough set of fragments, with sufficient overlap.

This mechanism can suffice by itself if enough overlapping fragments are used. This has been demonstrated by Minsky and Papert (1969) in their analysis of perceptrons, by showing that figures are determined uniquely by what they termed the figure's 3-vector spectra. Briefly, Minsky and Papert considered binary images of arbitrary black figures against a white background. They considered simple image fragments composed of triplets of black image points, and showed that the collection of fragments contained in a

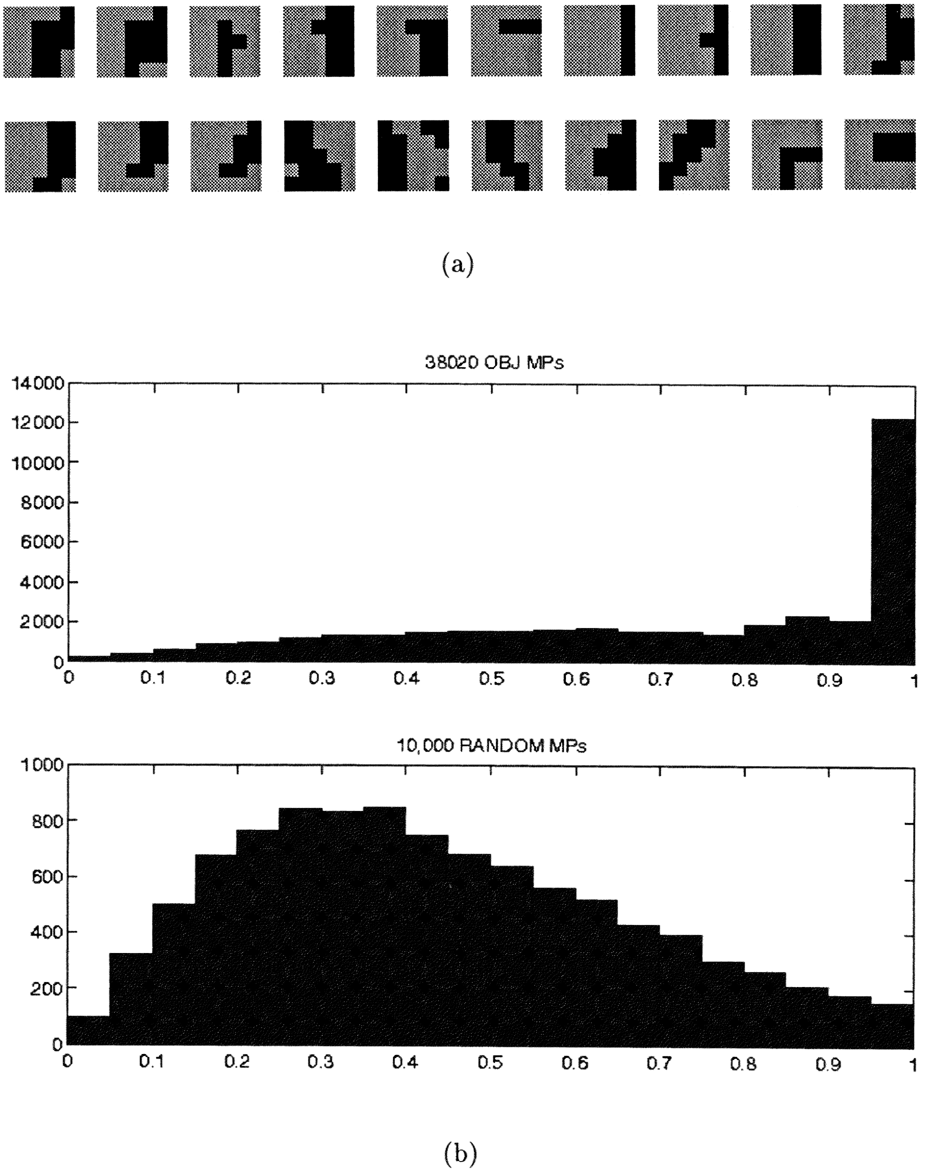

Fig. 6. (a) Examples of 5 × 5 micro-patterns obtained from natural objects. We consider the possibility that V1 detects micro-patterns of this type. (b) Histograms corresponding to the orientation/symmetry measures calculated for micro-patterns obtained from natural objects (top) and for randomly generated micro-patterns (bottom).

shape defines the shape uniquely. The extensive overlap between fragments is crucial for imposing the spatial configuration, and therefore schemes (such as the neocognitron by Fukushima, 1980) that do not exploit this property, are more prone to the rearrangement problem.

The use of primitive image fragments such as all triplets of figure-points has several disadvantages. In particular, the number of primitives is prohibitively large, and the scheme is relatively prone to occlusion. Our proposed scheme therefore employs a smaller number of more realistic fragments.

The use of image fragments to deal with shift and other forms of invariance requires some form of image segmentation, or coarse figure-ground separation, as a part of the recognition process. If the pattern we wish to recognize is embedded within a much larger and complex context, and if no image segmentation is performed, then, in addition to the fragments included in the target pattern, other fragments will be activated. This makes the detection of the target pattern more difficult and less reliable. To reduce the clutter of extraneous fragments, it is useful to restrict the region of analysis to the neighborhood of the shape being analyzed. Our simulations indicate, however, that only a rough form of image segmentation is required for this purpose. If the segmentation is imprecise, then some of the figure fragments will be missing, and some additional fragments, that belong in fact to the background rather than the shape, will be activated. A similar problem of missing fragments will also arise when a part of the shape is occluded by other objects in the scene. In either case, unless the system is required to make fine distinctions between shapes that have closely similar composition in terms of their fragments, the system will be able to tolerate the deletion or addition of image fragments. Unlike some other recognition schemes that rely on the detailed shape of the object's bounding contour, the fragment-based scheme is not required to delineate the pattern precisely, and it is inherently tolerant to clutter and occlusion.

In describing a shape by a collection of simpler fragments, it is natural to include fragments at different levels of resolution. For example, as can be seen in Fig. 2, in addition to partial fragments at high-resolution, there are also low-resolution fragments that give a coarse representation of the entire shape. The coarse global fragments are also useful in imposing the global configuration, and the smaller, high-resolution fragments capture the precise details.

To summarize, let us recapitulate briefly how shift invariance is obtained by the system. The main issue to examine is how a novel shape, learned at one location, will be recognized correctly at other locations. When presented at the initial location, the new shape will activate a set of fragment-detecting units at different levels of complexity. As mentioned earlier, the local fragment detectors are assumed to activate more global units that respond to the presence of their preferred fragment within a larger region of the visual field. The exact nature of the fragments incorporated in the system will depend on the history of the system, in particular, its prior experience with similar shapes. In natural cases, the shape to be recognized will often belong to a familiar class, and in this case, as we have seen in the simulations, the set of activated fragments is expected to provide a detailed and unique representation of the shape. When the shape is seen at a different location, the same set of (global) fragment detectors will be activated, leading to correct identification at the new location. If the shape is embedded at the new location within a larger context, then additional fragments will be activated. Detection of the target shape will now occur if this larger set of activated fragments contains all the shape's fragments as a subset. In this case identification will be less reliable. Reliable recognition at the new location can be obtained by segmentation or attentional processes that can restrict the gathering of relevant information to the neighborhood of the pattern to be analyzed.

If the target shape belongs to a novel set that is not well represented by existing fragments, then translation invariance will be more limited. The system will fail initially to make fine discriminations between the target shape and similar shapes. The degree of discriminability will depend on the already existing fragments: two shapes that have identical or similar representations in terms of existing fragments will be confusable. Following learning at a given location, the system will eventually store additional fragments that can deal with the new shapes. But since the new shapes were shown at one location only, the system will not have the machinery to generalize the newly acquired discriminations to other regions. Shift invariance for the new class of stimuli will develop gradually, as the system learns to extract additional fragments at multiple image locations. These properties of the fragment-based scheme agree well with the observed psychophysics of shift invariance, as discussed in Section 1.

3.2. Possible implications to V1

In the proposed scheme the system uses at each stage a repertoire of different patterns, starting from small micro-patterns and building more complex object-fragments. This view suggests that the initial analysis, at the level of the primary visual cortex, may not be restricted to the detection of a small number of primitives such as straight lines and edges, but a more elaborate repertoire of micro-patterns. Fig. 6(a) shows examples of small micro-patterns that re-occur in patches of natural images we have analyzed.

Although the micro-patterns extracted from natural images are not just straight edge and line fragments, they tend to show statistical preference for oriented local patterns. To examine this characteristic of natural micro-patterns, we used a measure of orientation vs. symmetry of micro-patterns. The distribution of the black pixels within the micro-pattern can be estimated by an ellipsoid at the center of the distribution, with its axes estimated by the pattern's second moments (specifically, by the square roots of the eigenvalues of the pattern's covariance matrix). The measure we used was calculated by the following formula:

$$\xi = 1 - e^{1-\frac{\lambda\max}{\lambda\min}},$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the eigenvalues of the covariance matrix. If the micro-pattern is completely symmetrical, then the eigenvalues are the same, and $\xi = 0$. If the micro-pattern is completely oriented, then the ratio of the eigenvalues becomes infinite, and $\xi = 1$. Note that the orientation measure does not change if black pixels are shifted inside of the micro-pattern, because the covariance matrix is invariant under translation of the pixels.

The orientation measure was calculated for the different 5×5 micro-patterns obtained from the set of natural objects and for randomly generated micro-patterns. Random binary matrices were taken from the discrete uniform distribution. The histograms corresponding to the calculated measures are shown in Fig. 6(b). It is clear from the figure that 5×5 micro-patterns obtained from the set of object images are significantly different in terms of their orientation distribution from those taken from randomly generated images. If the primary visual cortex contains in fact multiple units selective to many different types of micro-patterns that are prevalent in natural images, this population will show strong orientation preference of the type shown in Fig. 6. It will show a range of orientation preferences, with a large proportion exhibiting high orientation preference. Given the current state of knowledge, it is difficult to distinguish between the more standard view of V1 as encoding a restricted set of standard units, and the view suggested here of a much richer repertoire of different micro-patterns reflecting the occurrence of micro-patterns in natural images.

3.3. More general invariances

The scheme outlined above, where shapes are represented by their more primitive fragments, shares a number of basic properties with units in the visual system, including the gradual increase in receptive field size along the ventral visual pathway, and the selectivity to increasingly complex shapes. As expected from the scheme, the preferred stimuli of units along this pathway are usually characterized in terms of their preferred two-dimensional patterns (rather than, for instance, a three-dimensional primitive shape), and the preferred stimuli are usually image fragments of different complexity, rather than the shape of a complete object.

The scheme makes use of the high degree of parallelism that characterizes brain-like structures, using many units with different preferred shapes. It is primarily “memory based”, using sub-patterns seen in the past, and it uses simple computations that appear more biologically plausible than, for example, the use of internal shifting mechanism, or the use of abstract structural descriptions. The computations employed are fast, and do not require lengthy iterative calculations. These properties were illustrated so far using shift invariance as an example, and it is natural to examine the application of a similar approach to more general problems of invariant recognition.

A general property of the fragment-based approach is that it uses a large repertoire of partial generalizations built upon past experience, to deal effectively with novel patterns. Building upon its exposure to many visual stimulations, the system learns many instances of equivalence between sub-patterns. Based, for example, on the drift of patterns across the retina, it learns that a shape fragment at one location can be replaced by another shape at a neighboring location. This will lead in time to the creation of a unit that responds to this shape at multiple locations (Földiák, 1991; Parga & Rolls, 1998). This unit that responds to the same sub-pattern at different locations, now encodes a potential partial generalization, that can be used in the future in the context of new patterns. Over time the system will store a large number of partial shapes together with possible generalizations. When confronted with a novel stimulus, and assuming that the stimulus can be well encoded by exiting fragments, the system will be in a position to use its store of partial shapes and potential generalizations to identify the novel shape in its entirety as equivalent to a previously seen shape.

Similar principles can be applied to other aspects of invariant pattern perception. For example, the image of a three-dimensional object changes with the object’s pose, that is, the relative orientation in space between the object and the viewer. The human visual system is highly adept at compensating for the complex image transformations induce by changes in three-dimensional pose (Ullman, 1996). For objects in a familiar class such as a face or a car, remarkable generalization is obtained on the basis of a single image of a novel object (Moses, Ullman & Edelman, 1996). Here, too, the system can store partial generalizations based on past exposure to the class of objects in question. As in the case of acquiring shift invariance, it will learn from experience how different fragments of face images change their appearance as the face changes its three-dimensional orientation. A unit will respond, for example, to an image fragment depicting a partial view of the eyes region, an eyebrow, the hairline region and the like, at different spatial orientations. As before, these fragments do not correspond necessarily to well-defined face parts, such as an eye or a nose, but may represent any partial face fragment. As in the case of shift invariance, a high degree of spatial overlap between fragments will be used to impose the proper spatial arrangement. In addition to the use of overlapping image fragments, it is also possible to use non-overlapping fragments to specify a spatial relation in a qualitative manner, for example, that fragments from the eyes region must lie above a nose-fragment. We found that such units are useful in object classification tasks considered next, but we will not discuss them here in detail.

Another domain that can use similar principles is that of object classification. In object classification, the task is to assign an object, including novel objects, to a general class such as a face, a car, a house, and the like. Classification is a natural skill for humans, and is performed quickly and effortlessly even by young children. In fact, for humans, classification comes easier than individual object recognition. It is easier to recognize an object as a car, for instance, than to recognize a specific make, such as a Toyota Camry or a Honda Accord. For current computer vision systems, the opposite is true. Individual objects are easier to recognize in such systems, because particular object models can be specified with precision. In contrast, it is difficult to instruct a computer system to recognize a general class, since the shape of the individual objects within the class may vary considerably.

In the case of classification, equivalence can be defined between image fragments on the basis of a substitution relation: if the eye region in a face, for example, can be replaced by a different eye region, and still represent a coherent face, then the two image fragments representing the eye regions will be considered as two instances of a more general unit. This is similar to shift invariance, where two fragments representing the same shape at different locations were considered two instances of a more general unit. As in the case of shift invariance, the system will acquire a repertoire of fragments of different size and complexity, with equivalence, or replacement relations, that serve as a basis for generalization to novel shapes. In this scheme, a novel face, for instance, will be classified correctly even if it is quite different as a whole from all previously seen faces, provided that the eyes are sufficiently similar to the eyes of a face seen before, the mouth region to a similar region in a second face, and so on. Preliminary computational experiments, to be described elsewhere, with a classification system based on these principles showed good results in classifying a variety of face and car images.

3.4. Learning issues

A basic characteristic of the proposed approach that distinguishes it from most alternative approaches is that it depends on past experience and the learning of many partial examples rather than the application of some general rules. For example, shift invariance is obtained not by a general shift operation, but by the convergence of related image fragments from different locations. This raises important questions regarding learning from past examples. The learning task in this scheme can be divided into two main issues. First, how the system learns to extract useful shape fragments, and second, how it learns to associate related fragments and form more generalized units. We will not discuss these learning issues in detail, but only comment briefly on the acquisition of fragments and their associations.

With respect to image fragments, it appears that useful fragments can be extracted from image parts that are common to a number of different objects. If the same subimage appears as a common part in a number of different objects, then it is likely to be a useful building block for representing this set of objects. The common fragment is more likely to be useful if it is a structure of significant size, as then it is unlikely to have been generated accidentally. Even if just two images have a significant common structure, it may be worth while to extract the common part as a potentially useful fragment. This is because two unrelated images are statistically unlikely to have a significant common structure. If such a common part does exist, it is unlikely to be accidental, and it may represent a significant structure that will appear in novel objects as well. The learning process can therefore proceed on the basis of a small number of examples, rather than by seeking statistical regularities in a large population of stimuli, as in some statistical learning schemes.

The system needs also to associate together related fragments, for example, two similar fragments at different locations, or two fragments representing an object part at different three-dimensional orientations. This can be obtained by spatio-temporal continuity of patterns across the retina as objects move in the world, and as the gaze shifts across the scene. As indicated by the phenomenon of apparent motion, the visual system has a strong and probably innate tendency to associate together image regions on the basis of spatial and temporal contiguity. Several neural network models have shown how synaptic learning rules can implement such association between consecutive views (Földiák, 1991; Parga & Rolls, 1998).

For the purpose of object classification, the association between equivalent fragments can be based, as mentioned, on a replacement relation. If, within a pattern P, a fragment F_1 in one view can be replaced by a different fragment F_2 in another view, then it will be useful to store an association between the two fragments as a potential generalization in other contexts as well. As a specific illustration of this replacement relation, consider two views of the same face, one with a neutral expression, the other with a smiling expression. Most of the face pattern will remain unchanged, but some parts will be replaced, e.g. the neutral by a smiling mouth. The system can conclude that the face in this case is composed of an unchanged part, combined with another region that can take one of two (or more) forms. The learned association between the two fragments will be used in the future in the context of other patterns as well. That is, the two forms of the mouth regions will be treated as equivalent in novel face images, and this will allow the system to make useful generalizations.

In conclusion, the acquisition of invariant perception in this scheme depends on a continuous learning process. Based on its visual experience, the system will extract image fragments that are likely to be useful building blocks for more complex patterns. Based on certain relations between patterns, such as spatio-temporal contiguity and replacement, it will form associations between fragments and create higher-order, more abstract units that respond to a number of fragments that can be considered equivalent for the purpose of invariant perception.

4. Summary

Invariant perception is an achievement of biological visual systems that is difficult to replicate in artificial systems. We have outlined an approach to the computation of pattern invariance that appears more suitable for brain-like structures than alternative approaches. In this approach, invariance for complex patterns is based on a large number of stored relationships between more elementary image fragments. Invariant perception therefore depends on a continuous process of learning from visual experience. During this process, the system extracts and stores image fragments that are likely to be useful building blocks for representing a family of related shapes. Based on certain relationships between fragments, in particular spatio-temporal contiguity and replacement, the system creates higher-order units that respond to a number of different fragments. In the case of position invariance, for example, the shift of patterns across the retina will lead to the creation of units that respond to fragments representing the same shape at different locations. Such units can be viewed as encoding partial potential generalizations that can be used later in the context of novel shapes. The stored image fragments are constructed hierarchically, at different levels of complexity, and equivalence can be established between fragments at different levels.

We have illustrated how this approach can apply to shift invariance. Based on a limited number of fragment detectors that are replicated throughout the visual field, the system can recognize correctly a large number of more complex shapes in a shift-invariant manner. Shift invariance is therefore not a built-in mechanism, based, for example, on an internal shift circuitry, but an acquired skill built upon multiple partial examples. As a result, this scheme will share certain limitations of shift invariance with the human visual system. The generalization of shape recognition to new positions will not apply equally to all shapes. Novel shapes that can be represented well, with a high degree of discriminability, by existing fragments, will generalize well to new locations. For other shapes, that are not well captured by already existing fragments, shift invariance will be initially limited, and will require the learning of new fragments at multiple locations. We have also seen in the computational studies how the use of multiple overlapping fragments enforces the correct arrangement of the participating fragments. Enforcing the correct arrangement can also be aided by processes of image segmentation or by an attentional process that limits the region over which visual information is analyzed.

The scheme is consistent with a number of general properties of units in the visual system, including the gradual increase in receptive field size along the ventral visual pathway, and the selectivity to increasingly complex shapes. As expected from the scheme, intermediate units along the visual pathway have preference to a variety of different two-dimensional patterns, and the preference up to the level of anterior IT is often to partial image fragments rather than complete objects or object parts.

In human vision, it is likely that invariant perception involves more than a single mechanism. The scheme described above is therefore not supposed to account for all aspects of invariant perception and pattern classification. It seems particularly appropriate for aspects of recognition and classification that are immediate and can deal effectively with familiar classes of objects, including novel individuals within these classes. In such cases, classification and recognition are obtained within a short time (down to about 100–200 ms), leaving little or no time for complex iterative computations (Rolls, Tovee & Lee, 1991; Thorpe & Imbert, 1989). The computations in the proposed scheme can be accomplished for the most part in a single feed-forward sweep. Top–down connections that appear to play an important role in recognition and classification (Ullman, 1996) can play a useful role in the proposed scheme, but their role will not be analyzed here further.

The generalization capacity arises in this scheme not from mastering abstract rules or applying internal transformations, but from using the collective power of multiple specific examples that have partial overlap with the novel stimulus. To obtain such memory-based generalization, the scheme makes use of several characteristic properties of brain-like structures. It employs a high degree of parallelism, using a large collection of units with different preferred shapes. It relies on high degree of connectivity to associate and bring together related fragments. Finally, it relies on continuous modification and learning processes that use visual experience to form the appropriate associations. This results in a scheme that can use its past experience to deal effectively with novel stimuli and generalize to new situations.

References

Biederman, I. (1985). Human image understanding: recent research and a theory. *Computer Vision, Graphics, and Image Processing*, *32*, 29–73.

Biederman, I., & Cooper, E. E. (1991). Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, *20*, 585–595.

Bricolo, E., & Bülthoff, H. H. (1992). Translation-invariant features for recognition. *Perception*, *21* (Suppl. 2), 59.

Bricolo, E., & Bülthoff, H. H. (1993). Further evidence for viewer-centered representation. *Perception*, *22* (Suppl.), 105.

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *The Journal of Neuroscience*, *4* (8), 2051–2062.

Dill, M., Edelman, S. (1997). Translation invariance in object recognition and its relation to other visual transformations. Technical Report A.I. Memo No. 1610, MIT.

Dill, M., & Fahle, M. (1997). The role of visual field position in pattern-discrimination learning. *Proceedings of the Royal Society of London B*, *264*, 1031–1036.

Fiser, J., & Biederman, I. (1995). Size invariance in visual object priming of gray-scale images. *Perception*, *24*, 741–748.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*, 194–200.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193–202.

Gallant, J. L., Braun, J., & Essen, D. C. V. (1993). Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. *Science*, *259*, 100–103.

Gross, C. G. (1992). Representation of visual stimuli in inferior temporal cortex. *Philosophical Transactions of the Royal Society of London B*, *335*, 3–10.

Gross, C. G., & Mishkin, M. (1977). The neural basis of stimulus equivalence across retinal translation. In S. Harnad & R. Doty & J. Jaynes & L. Goldstein & G. Krauthamer (Eds.), *Lateralization in the Nervous System*, (pp. 109–122). New York: Academic Press.

Hubel, D. H., & Weisel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154.

Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, *73* (1), 218–226.

Jolicoeur, P. (1987). A size-congruency effect in memory for visual shape. *Memory and Cognition*, *15*, 531–543.

Le Cun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Back-propagation applied to handwritten zip code recognition. *Neural Computation*, *1*, 541–551.

Malach, R., Grill-Spector, K., Edelman, S., Itzchak, Y., Kushnir, T. (1998). Rapid shape adaptation reveals position and size invariance in the object-related lateral occipital (LO) complex. *Proceedings of the Fourth International Conference on Functional Brain Mapping of the Human Brain*.

Marr, D., & Hildreth, E. C. (1980). Theory of edge detection. *Proceedings of the Royal Society London B*, *207*, 187–217.

Minsky, M., & Papert, S. (1969). *Perceptrons. An introduction to computational geometry*, Cambridge, MA: MIT Press.

Moses, Y., Ullman, S., & Edelman, S. (1996). Generalization to novel images in upright and inverted faces. *Perception*, *25*, 443–461.

Nazir, T. A., & O'Regan, J. K. (1990). Some results on translation invariance in the human visual system. *Spatial Vision*, *5*, 81–100.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, *13* (11), 4700–4719.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1995). A multiscale routing circuit for forming size- and position-invariant object representations. *Journal of Computational Neuroscience*, *2*, 45–62.

Parga, N., & Rolls, E. (1998). Transform-invariant recognition by association in a recurrent network. *Neural Computation*, *10* (6), 1507–1525.

Perrett, D. E., Rolls, E. T., & Caan, W. (1982). Visual neurons responsive to faces in the monkey temporal cortex. *Experimental Brain Research*, *47*, 329–342.

Rolls, E. T., Tovee, M. J., & Lee, B. (1991). Temporal response properties of neurons in the macaque inferior temporal visual cortex. *European Journal of Neuroscience*, (Suppl. 4), 84.

Salinas, E., & Abbot, L. F. (1997). Invariant visual responses from attentional gain fields. *Journal of Neurophysiology*, *77*, 3267–3272.

Shiller, P. H. (1995). Effect of lesions in visual cortical area v4 on the recognition of transformed objects. *Nature*, *376*, 342–344.

Shiller, P. H., & Lee, K. (1991). The role of primate extra-striate area v4 in vision. *Science*, *251*, 1251–1253.

Soloviev, S. (1997). Shift-invariant recognition by the conjunction of basic invariant patterns. Master's thesis, The Weizmann Institute of Science, Rehovot, Israel.

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

Tanaka, K., Saito, H. A., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, *66*, 170–189.

Thorpe, S. J., & Imbert, M. (1989). Connectionism in perspective. In R. Pfeifer & Z. Schreter & F. Fogelman-Soulié & L. Steels (Eds.), *Biological constraints on connectionist modeling*, Amsterdam: Elsevier.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Ullman, S. (1996). *High-level vision: object recognition and visual cognition*, Cambridge, MA: MIT Press.