Available online at www.sciencedirect.com
SCIENCE DIRECT®
PERGAMON
Vision Research 43 (2003) 2073–2088
Vision Research
www.elsevier.com/locate/visres

An integrated network for invariant visual detection and recognition

Yali Amit [a,*], Massimo Mascaro [b]

[a] Department of Statistics, University of Chicago, Chicago, IL 60637, USA
[b] Dipartimento di Fisiologia Umana e Farmacologia, Universita di Roma, La Sapienza, Rome, Italy

Received 5 June 2002; received in revised form 5 December 2002

Abstract

We describe an architecture for invariant visual detection and recognition. Learning is performed in a single central module. The architecture makes use of a replica module consisting of copies of retinotopic layers of local features, with a particular design of inputs and outputs, that allows them to be primed either to attend to a particular location, or to attend to a particular object representation. In the former case the data at a selected location can be classified in the central module. In the latter case all instances of the selected object are detected in the field of view. The architecture is used to explain a number of psychophysical and physiological observations: object based attention, the different response time slopes of target detection among distractors, and observed attentional modulation of neuronal responses. We hypothesize that the organization of visual cortex in columns of neurons responding to the same feature at the same location may provide the copying architecture needed for translation invariance.

Keywords: Invariance; Attention; Object detection; Object recognition; Hebbian learning

1. Introduction

The visual system performs complex detection and recognition tasks *prior* to any eye movement, in the first several hundred milliseconds after stimulus presentation. These include rapid object recognition using covert attention to extrafoveal areas (Cheal & Marcus, 1997; Duncan, 1980; Eriksen & St. James, 1986; Henderson, 1991), and efficient target detection among distractors (Humphreys & Heinke, 1998; Treisman & Sharon, 1990; Wolfe, 2001).

A common characteristic of these phenomena is *translation invariance*. By translation invariant object recognition we mean the ability to direct *covert* attention to a particular extrafoveal location and classify the object at that location without eye movement. By translation invariant object detection we mean the ability to detect complex objects among very similar distractors, again without eye movement, at rates faster than implied by random serial processing (using covert attention) of all object locations. Indeed, some experiments on detection of a target among distractors show very little dependence on the number of distractors even for complex objects and distractors that are very similar (Horowitz & Wolfe, 1998; Wolfe, 2001). Extensive reviews can be found in Desimone and Duncan (1995), Kanwisher and Wojciulik (2000) and Chum and Wolfe (2000, Chap. 9).

Our goal is to provide a computational model for invariant detection and recognition starting from a simple conceptual constraint: object representations and object classifiers are learned and stored in one *central module*. Consequently training examples of objects need only be presented at one location in the field of view. Furthermore, assuming learning takes the form of synaptic modification, these modifications only occur in the central module, and need not be translated or transferred to different locations in the visual system that respond to different locations in the visual scene. In short, we are proposing a model able to reconcile translation invariant detection, recognition and learning, with the existence of a central module where classifiers and object representations are learned and stored.

The key ingredient in the model is a *replica* module with multiple copies of the local feature inputs. The particular design of the inputs and outputs from this layer, which are described in detail below, enables location based and object based selection, and hence

* Corresponding author.
E-mail addresses: amit@galton.uchicago.edu (Y. Amit), mascaro@titanus.roma1.infn.it (M. Mascaro).

0042-6989/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0042-6989(03)00306-7

translation invariant detection and recognition. It is of particular interest that the organization of visual cortex with columns of neurons responding to the same feature at the same location can be used as such a copying mechanism.

The use of multiple copies for location based attention has been proposed in the literature in several versions (see Humphreys & Heinke, 1998; Olshausen, Anderson, & Van Essen, 1993; Salinas & Abbott, 1997). We show that the *same* copying mechanism can provide the basis for object based attention as well, provided object representations have a very simple form: a list of feature–location pairs. Each pair determines a feature that is present with high probability on the object at a prescribed location given that the object is at approximately the reference scale. If a sufficient number of such features is present at the correct location relative to a hypothesized center, this center becomes a candidate location for the object.

Using relatively simple features consistent with those known to be computed in the retinotopic layers V1, V2 and V4, we can actually produce such object representations, for a wide family of objects, yielding robust detectors on *real* gray level images. Furthermore using the same features we are able to produce classifiers with competitive performance. Training the models requires a very simple form of local Hebbian learning. Thus the proposed model not only provides a conceptual framework for dealing with translation invariance, but leads to a real working algorithm for computational vision. In this context we emphasize that the way the detector and classifier are constructed from local features takes into account not only translation invariance but also invariance to a range of linear and non-linear variations in the object populations, and is robust to occlusion and clutter. The same scheme can be extended to account for rotation and invariance over large scale ranges, although this is beyond the scope of the paper.

The proposed architecture offers a simple framework for interpreting the above mentioned psychophysical experiments on object detection and recognition with covert attention (Eriksen & St. James, 1986; Henderson, 1991; Horowitz & Wolfe, 1998; Wolfe, 2001) as well as a number of electrophysiological studies on modulation of neural responses as a function of attention (Chelazzi, Miller, Duncan, & Desimone, 1993; Connor, Gallant, Preddie, & Van Essen, 1996; McAdams & Maunsell, 2000; Moran & Desimone, 1985; Treue & Martinez, 1999). In particular location based and object based attention emerge as very similar processes, whereby different sections or slices of the replica module are primed.

Some predictions emerge as well. For example we expect that neurons in a single vertical column in area V4 and perhaps V2 and V1, corresponding to the same location and feature, would have differential responses to the preferred stimulus in their common receptive field. The factor determining this differentiation will either be the attended location, or the model of the target object that is to be detected. Furthermore, when object based attention is in effect, we expect increased activity among neurons throughout a retinotopic layer such as V4, even in the absence of the stimulus. This activity corresponds to the priming of the object model at all possible shifts. This is discussed in further detail in Section 6.5.

1.1. Relation to other work

In Olshausen et al. (1993) a shifting mechanism is proposed as a means for location based attention. Shifting is achieved through direct control of synaptic connections between the input layer and the central module. Such control mechanisms do not appear very plausible biologically. Our model makes use of multiple copies of the input features yielding a simple shifting mechanism similar to that proposed in Humphreys and Heinke (1998). Similar ideas can also be found in Salinas and Abbott (1997) who use a multiplicative gain that is a function of the distance between the preferred attentional locus of a neuron and the attended location. The limitation of their model is that learning of the object needs to be performed everywhere in the scene. We overcome this problem using a summation layer that integrates all the activities in the replica module. The data in this summation layer is subsequently classified. Our approach differs from models of the type presented in Deco (2000), where a retinotopic array of recognition neurons is required to deal with translation invariance. This would imply again that learning has to occur at each location, and that classification modules are present at each location.

As in Riesenhuber and Poggio (1999) we make use of maximization or 'spreading' (similar to that of the complex cell) as a mechanism to achieve some degree of robustness to geometric variability. This idea can also be found in Fukushima and Miyake (1982). However in Riesenhuber and Poggio (1999) the underlying premise is that each unit in the central module corresponds to a particular feature and is activated if *any* unit in a retinotopic array of detectors for this feature is activated. Thus any information on spatial organization is lost. One of the motivations for such spatially coarse models is to overcome the 'spatial binding' problem of different features at different locations. However in real complex scenes with multiple objects sharing many of the same features this form of processing is possible only if sufficiently complex features are employed. One would essentially need features at the level of complexity of the objects themselves, detected in retinotopic arrays, leading to a combinatorial explosion of the number of units required. Indeed the need to preserve spatial informa-

tion in the data being processed is the motivation behind the shifter mechanisms used in Olshausen et al. (1993), Salinas and Abbott (1997) and Humphreys and Heinke (1998). In the model below the spatial binding problem is resolved using the replica module.

In Olshausen et al. (1993) the subimage to be shifted into the central module is selected using a bottom-up mechanism identifying salient regions in the image. This idea of bottom-up selection using saliency maps has been further developed in Itti, Koch, and Niebur (1998). The form of selection studied below is primarily guided by top-down flow of information in the form of simple object representations. However clearly both mechanisms are useful and can be integrated into the architecture.

1.2. Paper organization

In Section 2 we describe the entire architecture of the model allowing for translation invariant detection and recognition. In Section 3 we discuss which features could be used as the basic retinotopic input into the system, in particular the payoff between feature complexity which yields more power, combinatorics, and spatial accuracy. In Section 4 we outline the learning mechanisms used for creating the classifiers and object representations, in Section 5 we provide an illustration on synthetic scenes produced from randomly deformed and randomly placed characters as well as on real images where the task is to detect faces. In Section 6 we try to explain a variety of experimental results reported in the literature in the framework of the proposed architecture, and outline some predictions.

2. Modeling translation invariance

We introduce the notion a reference grid G of limited size, say 20×20, and a family of binary local image features $f = 1, \ldots, N$. The central module, where object representations and classifiers are learned and stored, receives inputs through a system W of units $W_{f,z}$ corresponding to all feature–location pairs (z, f), $z \in G$, $f = 1, \ldots, N$. Object representations and classifiers are defined in terms of these feature–location pairs, as described next. The entire image grid corresponding to the full field of view is denoted L, with width and height of several hundred pixels.

2.1. Translation invariant object recognition

It is a well accepted fact that the visual system can direct *covert* attention to extrafoveal areas, enabling what amounts to translation invariant object recognition. Given that learning occurs only in the central module, there must be a mechanism for shifting the data in the attended area into this module for processing. This is the rationale behind the shifter circuits proposed in Olshausen et al. (1993). The authors suggest an implementation using short term modifications of *synaptic* connections between the retinotopic layer of features and a reference grid of features. This requires a complex mechanism for providing direct input to synapses as opposed to neurons. A simple alternative is to produce a physical copy of each shifted window. To simplify the description of the computation this is first described as a rather artificial stack of copies. Then in Section 2.3 we show how this stack can be arranged in a more plausible manner.

A retinotopic layer F detects instances of each feature f everywhere in the field of view. For each f let $F_{f,x}$ denote the unit responding to feature f at location $x \in L$. For each x define a copy layer (the size of the reference grid) U^x of units $U^x_{f,z}$, $f = 1, \ldots, N$, $z \in G$. The number of copies is given by the number of locations in L. A unit $U^x_{f,z}$ receives input from $F_{f,x+z}$, so that U^x is a copy of the data in the shifted window $x + G$. For a given pair (f, z), all units $U^x_{f,z}$, $x \in L$, feed into $W_{f,z}$ (the input units of the central module) which is activated if *any* of the units $U^x_{f,z}$, $x \in L$, is on.

A retinotopic layer S codes for selected locations. If S_{x_0} is activated, attention is focused on location x_0 by enhancing the overall input to the units in U^{x_0} and inhibiting the input to all units in other U^x layers. Each unit $W_{f,z}$ then 'sees' only the activity in $U^{x_0}_{f,z}$. This is summarized as

$$W_{f,z} = \sum_{x \in L} S_x \cdot U^x_{f,z}. \tag{1}$$

The input to the central module is thus restricted to the data copied from the window $x_0 + G$. It is classified using connections between W to a system of units A in which random subpopulations code for each class (see Section 4). Training of the classifier thus requires only the modification of the synaptic connections between W and A.

The idea is illustrated in Fig. 1. Note that overlapping regions get copied so that the configuration of features within each region is preserved. In the figure we use only one local feature—indicated by the darker points—as opposed to the real model where many types of local features are computed at each location.

2.2. Translation invariant object detection

Experiments on detection of a target among multiple distractors as described in Horowitz and Wolfe (1998) and Wolfe (2001) demonstrate that in certain situations the response time for detection of complex objects among complex targets is virtually flat as a function of the number of distractors. The rich family of objects and distractors employed in these experiments suggests that

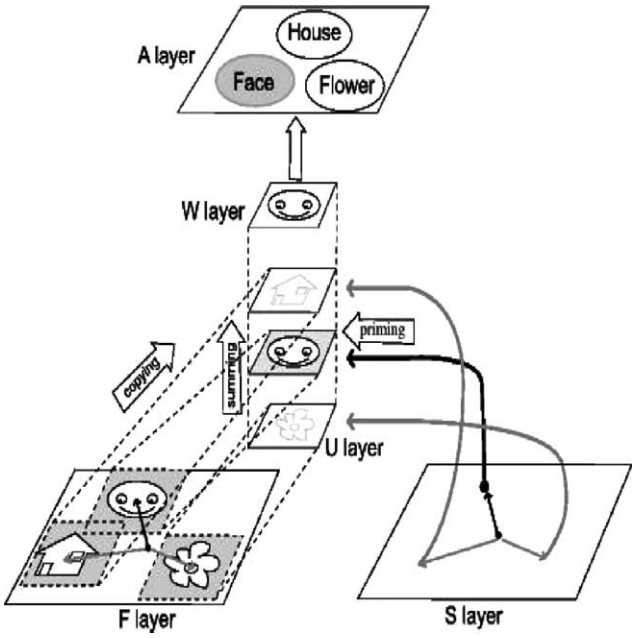

Fig. 1. Location based attention. A location x selected in the S layer primes the copy layer U^x and suppresses the input to all other U layers. All activity in the U layers is summed into W (see Eq. (1)), and subsequently classified through the $W \to A$ connections.

detection is not based solely on one local feature hard wired retinotopically in one of the layers of the visual system. It would be hard to reduce the efficient search for upside-down animals among upright ones (Wolfe, 2001), or for complex shapes among similar distractors, to the detection of a single pre-existing feature that flags the target against the distractor. An alternative explanation is that mechanisms exist whereby models are conveyed top-down and prime the system to respond to particular *configurations*, defined in terms of the hard wired features in F, independently at *each* location. Priming simply involves enhancing the input to particular neurons in a manner similar to the location based priming described above. Conceptually we need a dual shifting mechanism: shifting of a model to each location, together with a mechanism for comparing the model to the data.

The mechanism for comparing an object model to data must be very simple in nature, since it occurs at all locations simultaneously. Moreover it occurs very fast (response times can be on the order of several hundred milliseconds), and is not too accurate. After all mistakes are often made in detection tasks. As we will see in Sections 5.1 and 6.1, the false positive rate of the detector, which depends on how well the target model fits the distractors, can be used to explain the response time in detection tasks. In natural scenes in addition to specific distractors there is generic background clutter, and the false positive rate there is largely determined by the statistics of the individual features used to construct the models.

We assume the object models are represented in terms of a collection of feature location pairs in W. Given the object is present at reference scale in the reference grid, a collection (f_j, z_j), $j = 1, \ldots, n$, of feature–location pairs is found in training to be of high probability. Assume first that all object models have the same number of features. Define x as a candidate location of the object if

$$\sum_{j=1}^{n} F_{f_j, x+z_j} \geqslant \rho \tag{2}$$

for some fixed threshold ρ. This is a simple thresholded template match. The template consists of the list (f_j, z_j), $j = 1, \ldots, n$, and is compared with the input data around location x.

Connections from W to the system of U layers are defined as follows. Each unit (f, z) in W feeds into every unit $U^x_{f,z}$ for all $x \in L$. When the object model (f_j, z_j), $j = 1, \ldots, n$, is evoked in W all units $U^x_{f_j, z_j}$, $x \in L$, receive input, i.e. are primed, from their counterparts in W, and all other units are suppressed. Only those units in U^x that are both primed and receive input from the F layer are activated. Finally each unit x in the retinotopic layer S sums the input from all units in U^x, and is activated only if the sum is above ρ. Thus the input to a unit S_x in S is defined as

$$I(S_x) = \sum_{f,z} U^x_{f,z} \cdot W_{f,z} = \sum_{j=1}^{n} F_{f_j, x+z_j}, \tag{3}$$

and S_x is on if $I(S_x) > \rho$. Those units active in S will correspond to candidate locations of the object defined by Eq. (2). If more than one unit is activated a competition using inhibitory connections can yield one selected location.

Models for different objects may have different numbers of features so that ρ needs to change as a function of n. This would require some mechanism for modulating the baseline activity in S depending on the number of features activated in W. Note that the S layer could also serve to flag those locations selected for attention using a bottom-up saliency map as proposed in Itti et al. (1998), for example in the absence of a top-down guided detection task.

The architecture is shown in Fig. 2, where for illustration purposes, the overlap between data and model is exact. In Section 3.1 we describe how to obtain a much greater degree of flexibility using a spreading operation, so that the detection process is robust to a wide range of deformations of the object.

2.3. The full architecture

The full architecture including the direction of the synaptic connections is summarized in Fig. 3. The abstract representations of the different classes are located in A, in terms of a distributed population code. For

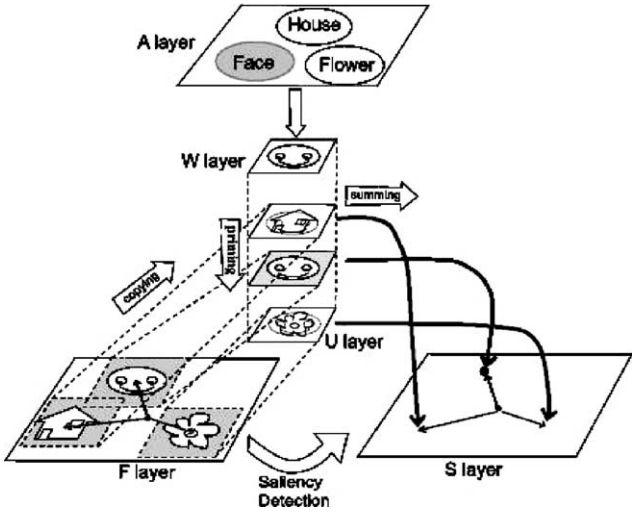

Fig. 2. Object based attention. When an object class (face) is evoked in A the feature–location pairs of the object model are turned on in W and prime the corresponding feature–location pairs in *all* the U layers (dim copies of the face). Only in a U^x layer where this top-down input is reinforced by a similar bottom-up input from F is there output to the corresponding unit x in S (see Eq. (3)).

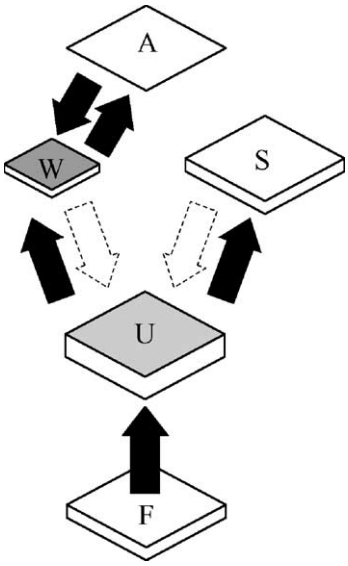

Fig. 3. A summary of the architecture. The arrows indicate the directions of the synaptic connections. Dashed arrows are represent connections involved in priming.

recognition, the selected location is determined in S either through top-down selection or through bottom-up selection using a saliency map. This primes a particular U^x layer. Only the data from U^x arrives at W and is then classified using connections from W to A, causing the appropriate subpopulation to be activated. For detection the desired class is excited in A evoking the feature representation in W, subsequently priming copies of the features in each of the U layers for comparison to the

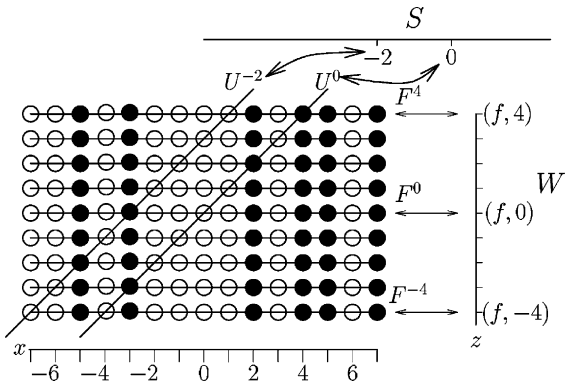

Fig. 4. The replica module: producing the U layers with copies of F. Multiple retinotopic copies of F layers are shown indexed by $z \in G$. A diagonal line passing through x at F^0 corresponds to the layer U^x, and all units $F^z_{f,x+z}$, $z \in G$, $f = 1, \ldots, N$, along this line have connections feeding into and out of the unit $x \in S$. For fixed (f, z) all units $F^z_{f,x}$, $x \in L$, have connections to and from the corresponding unit $W_{f,z} \in W$. Assuming 40 types of features, a 20×20 reference grid, and a 100×100 image grid, the total number of units in the replica module is $400 \times 40 \times 10{,}000 = 1.6 \times 10^8$.

incoming data. Those U layers with sufficient activity activate the corresponding location in S.

For expository purposes we have distinguished between the F layer where the features are extracted and the U layers where they are copied. This copying mechanism can also be implemented using multiple copies of retinotopic F layers, one for each $z \in G$, which we denote the *replica module*. This is illustrated in Fig. 4. For each location $z \in G$ define a retinotopic layer F^z detecting all the features everywhere in L. A unit $W_{f,z}$ feeds into and receives input *from all* units in $F^z_{f,x}$, $x \in L$. On the other hand each unit $x \in S$ receives input and feeds into all units $F^z_{f,x+z}$, $z \in G$, $f = 1, \ldots, N$ (diagonal lines). It is not hard to see that this produces the exact same connections as the architecture of the U layers. However now units responding to the same feature at the same location are placed in a *vertical column*. The receptive fields are the same as well as their preferred feature, as observed in numerous electrophysiological recordings. We return to this point in Section 6.2.1 in the context of the experiments reported in Connor et al. (1996). Note that the number of required copies of the basic retinotopic input layer F is *only* the size of the reference grid (400 in our model).

2.3.1. Interaction between detection and recognition

This simple architecture implements translation invariant recognition and detection as processes involving priming of certain sections of the retinotopic F layers and summation on other sections. The two processes can easily *interact*. A particular detection task initiated by activating a subpopulation in the A layer leads to the selection of a particular location in $x \in S$ where the data has a sufficiently good match to the primed object

model. The activation in S can in turn initiate the selection of the input data corresponding to that location. The W layer then sees only the data around x which is classified through the connections with the A layer. This can be used not only to verify the correctness of the selected location but to provide the class of the object at that location.

In the following section we deal with invariance to scale and other deformations and the combinatorics involved in the proposed system of replicates.

3. Invariance, accuracy and combinatorics

Object detection and recognition are both highly robust to a range of linear and non-linear deformations of the objects. It is therefore important to incorporate such invariance in the architecture. On the other hand in real scenes there are plenty of distractors in the form of other objects or parts of objects that the biological system is able to ignore. Statistically speaking invariance is intended to reduce false negatives (increase sensitivity), but this must not come at the expense of too many false positives (loss in specificity).

Here we propose using models and classifiers based on complex local features, with a degree of slack in their prescribed position on the reference grid. The greater the slack the greater the sensitivity, the higher the complexity the greater the specificity. However the cost that accrues with increased complexity is in the number of feature types required to model all objects, leading to a tradeoff between accuracy and combinatorics.

3.1. ORing—a mechanism for invariance

The feature–location pairs used in the object models are supposed to have a high probability given that the object is present in the reference grid at the reference scale. There are significant variations within an object class even at a fixed scale, moreover instances of the object will never be present at precisely the same scale, even during training. High probability local features over the entire object class will be found only if the features themselves are robust to local variations in the object.

Assume that the initial input into the system are the elementary oriented edges known to characterize the responses of simple neurons in V1. A 'simple' edge detected at a particular location on an object will not be detected at the same location if the object is slightly scaled, or deformed in that area. The solution is to perform an 'ORing' operation at each location. Specifically we define a new layer of neurons, analogous to 'complex' neurons in V1, that respond if there is an edge of a particular orientation *anywhere* in the neighborhood of a location i.e. anywhere in the neuron's receptive field. The ORing can also be viewed as a 'spreading' operation in which each detected edge is spread to a neighborhood of the original location. Such units were used in Fukushima and Wake (1991) and Amit and Geman (1999). The analogue for continuous valued responses is the 'MAX' operation proposed in Riesenhuber and Poggio (1999). It is important to note that even when spread edges are used the object models and the classifiers are based on the spatial layout of these new features in the reference grid.

3.2. Complexity—a mechanism for specificity

In principle the 'spread' edges could constitute the local features feeding into the recognition and detection system. They allow for a degree of local invariance, and would yield high probability local features on the object. However from our experiments with real gray level images, edges with sufficient contrast invariance are quite common, and their density only increases after spreading. They will have very little discriminating power against the background, and the resulting detector expressed through Eq. (2) would yield numerous false positives. The only way to avoid false positives is to ensure that the background density of the features is low.

The solution we propose, and which appears to be chosen by the biological system as well, is to define more complex features. The term complex here refers to the structure of the feature. The features are defined as functions of the original edges, specifically local conjunctions. The simplest form is a pairwise conjunction of two edges of particular orientation, where the second is a 'spread' edge constrained to lie at some angle relative to the first. Using a 'spread' edge in the conjunction introduces some robustness in the definition of the feature. In Fig. 5 we illustrate 15 of the 40 such pairs that can be thought of as coarse *curvature* detectors. The density of such features in real images is much lower than that of edges. On the other hand it is still possible to find high probability features on a given object. Thus each feature has more discriminating power between

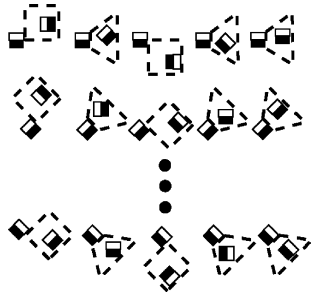

Fig. 5. Complex features: flexible pairwise edge conjunctions. The second edge can be anywhere in the designated region relative to the first.

object and background (see Amit, 2002; Amit & Geman, 1999). Such local features are also consistent with experimental findings on responses of neurons in V2 (see Hedgé & Van Essen, 2000). These 40 features are employed in the experiments below.

3.3. Accuracy vs. combinatorics tradeoff

More complex features can be defined using more edges in the conjunctions. Typically the density on background falls much faster than the probability of the best features on object. The result is that more complex features provide better inputs in terms of the detection algorithm. Invariance obtained through spreading is achieved with a relatively low price in false positives. On the other hand the number of features needed to represent all objects increases rapidly with their complexity. If the features are hard wired in multiple copies of retinotopic arrays this poses a combinatoric problem.

One remedy is to use lower spatial resolution. Indeed after a feature is spread, there is significant loss in location accuracy, and there is little loss of information if the original grid on which the features are detected is subsampled. The more complex the features, the more they can be spread and still yield high ratios of object to background probabilities. The more the features are spread, the greater the reduction in spatial resolution. In V2 features are more complex than in V1, but the spatial resolution is reduced, and V4 is an additional step in that direction.

In the specific model described here there are 40 features, and we assume the field of view at the lower resolution used for the complex features is 100×100. Thus L contains 10,000 points and the reference grid contains 400 points. The replica module will then contain $400 \times 40 \times 10{,}000 = 1.6 \times 10^8$ units. Downsampling by a factor of 2 for even more complex features would save a factor of 4. On the other hand increasing the feature complexity by adding an additional edge to each pairwise conjunction, could multiply the number of features by several orders of magnitude. It would be impossible to counter this increase in the number of features with an equivalent decrease in spatial resolution.

In the architecture presented here the F layers represent the locations of the edge-conjunctions after 'spreading' and subsampling. The processing prior to the F layers involves detection of the edges and of the edge conjunctions at the original resolution. This is illustrated in Fig. 6.

The tradeoff between feature complexity, invariance, accuracy and combinatorics is of great importance and little is understood apart from some empirical findings, see for example Amit (2002).

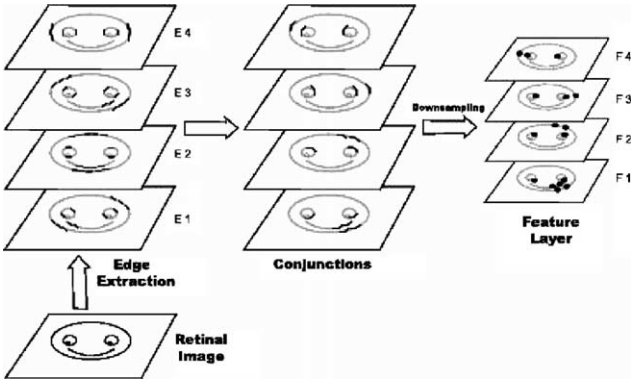

Fig. 6. From retinal image, to edges, to edge conjunctions. These are spread and subsampled to produce the data in F.

3.4. Why binary features?

Typically, orientation selective units in V1 are modeled as Gabor filters with continuous output (see Deco, 2000; Riesenhuber & Poggio, 1999; Wilkinson, Wilson, & Habak, 1998; Wiskott, Fellous, Kruger, & von der Marlsburg, 1997). However, the visual system is highly robust to contrast variations even in rather small neighborhoods of the same scene. It is plausible that the initial fast processing needs only coarse output of the form: is the response higher or lower than some rather conservative threshold? Hence our use of binary oriented edges. Furthermore the definition of more complex features (conjunctions) in terms of the binary edges is very simple, as well as the ORing operation used to obtain invariance. A final advantage of binary features is the simplicity of parameter estimation. For continuous features it is necessary to model the continuous distribution and estimate the relevant parameters. As we will see below estimation of simple probabilities is achieved using the simplest forms of Hebbian learning.

Despite the significant reduction in information caused by binarization, we find that the resulting recognition and detection algorithms perform well on real data. It is possible that the continuous valued data is only needed after foveation, when more intense processing is performed in the foveal region.

4. Training

All units in this architecture are binary computing the weighted sum of their inputs and comparing it to a fixed threshold (the same for all neurons). Each synapse has an internal state variable Y that is allowed to vary between two reflecting barriers. The actual efficacy is $J = H(Y)$ where H is a ramp like function varying between 0 and some maximal value $J_{\max}$. Learning affects the internal state of the synapse, in a Hebbian fashion

$$\Delta Y_{uv} = \alpha uv - \beta u(1-v), \tag{4}$$

where Y_{uv} is the state of the synapse that connects units u (presynaptic) and v (postsynaptic). The units u and v can only assume the values 0 and 1 so that synaptic modification occurs only when the presynaptic unit is active. The result is an increase by amount α (potentiation) if the postsynaptic unit is also active and a decrease by amount β (depression) if not.

The system of units A, introduced above, is composed of a large number of neurons with randomly selected subpopulations coding for each object class. The subpopulations denoted A_c are selected before learning. It is assumed that each unit in W receives input from a random subset of the units in A. The inputs are randomly distributed among the subpopulations A_c. Similarly each unit in A receives inputs from a randomly selected subset of W. During training, when an image from class c is presented to the network, the detected features activate the corresponding units in W. At the same time the entire subpopulation A_c is activated in A. Subsequently, both synaptic connections from A to W and those from W to A are modified.

4.1. Learning object models

For object models the $A \to W$ synapses are modified. All synapses for which a presynaptic unit belongs to A_c will be modified. A positive change (potentiation) of magnitude α occurs if the postsynaptic unit in W is activated by the example. Otherwise a negative change (depression) of magnitude β occurs. Let $p_{(f,z),c}$ be the probability that the feature f is present in an example drawn from class c at location z on the reference grid. Then all synapses that connect a unit in A_c to the unit $W_{f,z}$ will undergo, on average, $Np_{(f,z),c}$ potentiations and $N(1-p_{(f,z),c})$ depressions after N examples of class c are shown. On average the net change in the internal synaptic state will thus be $\langle \Delta Y \rangle = N(\alpha p_{(f,z),c} - \beta(1-p_{(f,z),c}))$. If N is large the synapse will be active only if

$$\alpha p_{(f,z),c} - \beta(1-p_{(f,z),c}) > 0 \to p_{f,c} > \frac{\beta}{\alpha+\beta}.$$

This leads to the selection of those feature–location pairs with *on object* probability above a given level.

4.2. Learning to recognize

For classification we modify the $W \to A$ synapses. We emphasize that each neuron in A receives input from a *random* subset of the neurons in W. The problem now is more complicated, since the units in A are postsynaptic and are required to identify the class of the object registered in W, by activating the correct subpopulation A_c. This problem has been studied in detail in Amit and Mascaro (2001), where it was shown that even a simple layer of perceptrons, when aggregated in populations, can achieve good results on rather difficult problems.

The simple Hebbian learning rule of Eq. (4) is insufficient if objects have variable feature statistics, as do real objects. The modification suggested in Amit and Mascaro (2001), in its simplest version, has the coefficient α turn to 0 if the total current input to the unit, which is a function of the current training sample and the state of all afferent synapses, is above some threshold, greater than that unit's firing threshold. This may be viewed as modulating potentiation by the firing rate of the post-synaptic neuron. This solution has proved to be stable and enables the use of fixed thresholds for all neurons in A. There is no need to adjust these thresholds or to perform a global normalization of synaptic weights. We observe that the potentiated synapses typically come from units $W_{f,z}$ with the highest ratio

$$\frac{p_{(f,z),c}}{q_{(f,z),c}}, \tag{5}$$

where $q_{(z,f),c}$ is the probability of (f,z) on the population of training samples not from class c. These are the most informative features for discriminating between class c and the rest of the world.

At the end of training, each of the units in the set A_c is a perceptron classifying class c against the rest of the world, in terms of input from a random subset of feature–location pairs. Classification is then obtained by a 'vote' among the different A_c populations. The vote is implemented through attractor dynamics, resulting from recurrent connections within A. This will tend to enhance the activity in the subset A_c with highest input and suppress the activity in the other subsets. Since each neuron in A receives input from a different subset of W, the final classifier is an aggregation of multiple randomized perceptrons, and can produce complex decision boundaries, despite the simple linear form of each of the individual perceptrons.

5. Computational experiments

5.1. Synthetic data

We present a simple example aimed at showing the effectiveness of the proposed architecture on synthetic scenes with multiple characters randomly placed and perturbed. Some results on classifying handwritten characters with the recognition architecture are described in Amit and Mascaro (2001). We use a set of 26 LaTeX characters, shown in the left panel of Fig. 7. Natural variability has been implemented by adding some degree of deformation as shown in the right panel of Fig. 7. We sample uniformly from a range of rotations of $\pm 15°$, and in log-scale independently in the two coordinates, between $\log(0.8)$ and $\log(1.2)$.

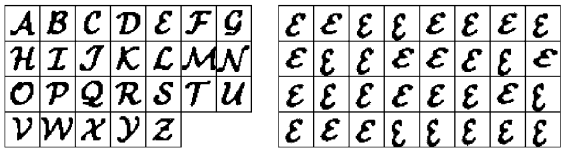

Fig. 7. Left: The 26 prototypes used in the experiment. Right: The training set used for the $\mathscr{E}$. An analogous set was used for each of the 26 symbols.

The reference grid is 20×20. Using the 40 edge conjunctions defined in Section 3.1, the W layer contains $40 \times 20 \times 20 = 16{,}000$ units. The feature extraction process is illustrated in Fig. 6.

Layer A is divided in subpopulations, one for each of the 26 classes. Twenty units are allocated for each class, and for simplicity they are taken to be non-overlapping. Thirty two examples for every one of the 26 classes are presented to the system. We train both the connections leading from W to A for classification, and the connections from A to W for object representations. The feature–location pairs evoked in W by class $\mathscr{E}$ are shown in Fig. 8. The features are shown in the location they are expected on the reference grid. This is the representation of class $\mathscr{E}$ used for detection.

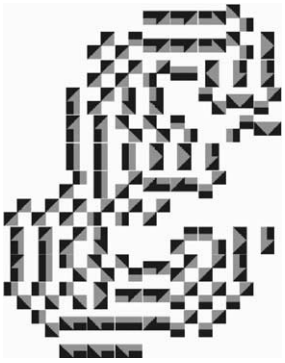

Fig. 8. The collection of edge conjunctions identified as having high probability on the $\mathscr{E}$ at their expected location on the reference grid. These feature–location pairs turn on in W when the class $\mathscr{E}$ is evoked in A.

After learning, the system is tested in a visual detection and recognition task on a scene containing the $\mathscr{E}$ and 20 randomly placed distractors of some other class. In these scenes, in addition to the random affine perturbations used in training, random Fourier coefficients at very low frequencies are used to define smooth nonlinear deformations, to further deform the objects. We also add random background clutter composed of small parts of characters.

The system locates candidate locations, using the top-down priming mechanism described in Section 2.2. On a serial computer this amounts to detecting all 40 types of edge conjunctions features in the image, spreading, recording their location on a coarse grid, and finally matching the model at every location on the coarse grid. The locations above some threshold are flagged. One can then process the location with highest match to the template or simply select a detected candidate location at random. Only the data in the window around the candidate location is 'seen' by W for further classification. The final part consists in updating the state of the A layer based on the input from W. A vote among the units in the subsets A_c determines the class label. If this is different from the target class $\mathscr{E}$ the location is discarded. For illustration purposes we also show the classified false positives in the image.

Three types of distractors are shown in Fig. 9. The first, $\mathscr{N}$, is very different in shape from $\mathscr{E}$, hardly any false positives are detected by the top-down detection mechanism, and it remains to verify that the unique candidate location is indeed an $\mathscr{E}$ through the classification mechanism. The second distractor is $\mathscr{B}$ where about half the distractors are detected, and subsequently need to be classified in order to be discarded. Finally the third distractor is the $\mathscr{S}$ where almost all distractors are detected as candidate locations and all must be processed through the classifier. Note that the response time curve in the first case would be flat, the response time curve in the third case would be steep—the single item classification time multiplied by the number of distractors.

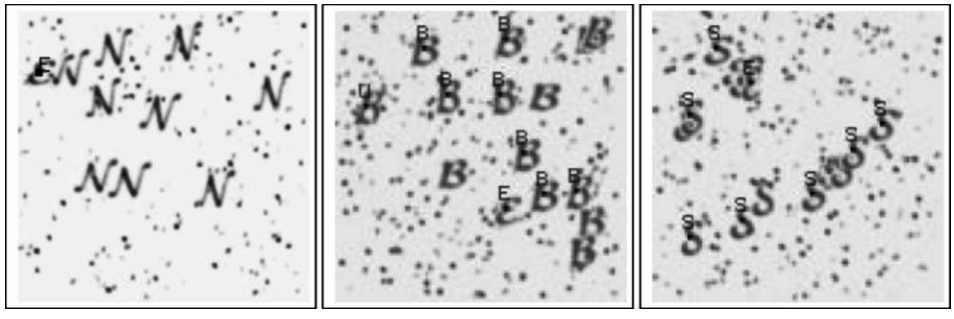

Fig. 9. Examples of detections and classifications. All detections of the $\mathscr{E}$ model, for three types of distractors, are marked with black square dots. The final classification of each detection is shown as the character above the dot. In addition to random distractors some random clutter has been added.

The second case would be intermediate. We discuss this further in Section 6.1. Robustness to clutter is quite evident from these scenes. Both the detection and the classification are at times performed with objects very close to or occluding the target, in addition small fragments of characters have been added as background clutter.

5.2. Faces

Face detection has been a relative success in computational vision. Recent successful and very efficient algorithms are found in Fleuret and Geman (2001) and Viola and Jones (2002). In both cases detection involves a sequence of simple perceptron type classifiers for object vs. background, defined in terms of a family of local binary features. Certain aspects of these algorithms could be implemented using the architecture described here. In the present context face detection is of interest because we are dealing with real objects in real scenes. The background is typically complex, including other objects, object parts, textures, etc. The proposed model does produce false positives but their numbers are not very large, and a subsequent classification of face vs. non-face reduces them even further.

The training procedure for the face detector is identical to that of the character detectors above, using the same family of 40 features, and a 20×20 reference grid. This is very similar to the detector described in Amit (2000). Here however the detected region is passed on to a classifier, which decides face or non-face. In other words the face model is evoked in A, the appropriate units in the replica module are primed and the detected locations are computed in S. These locations are then visited sequentially, each one serving as a selected location. Through priming in the replica layer (this time for location) the feature data around the selected location is transfered to W for classification through the $W \to A$ connections.

The A layer contains 100 neurons, 50 for face and 50 for non-face. The detector is trained only on 300 positive examples of faces from the Olivetti data base, identifying high probability features on this population. The classifier is trained on examples of faces and examples of non-faces, 300 of each, and serves as a second step to prune out false positives from the detection. In Fig. 10 we show some results. The squares surround a detection that passes the classification test. The black dots show detected locations that were rejected by the classifier.

A more sophisticated implementation would train faces against false positives of the face detector. Indeed it would be of interest to see how such a classifier could emerge from the dynamics of the proposed architecture. This however is beyond the scope of the current work.

6. Analysis of experimental findings

6.1. Object based attention

The concept of object based attention is becoming widely used as a model for many visual phenomena, see for example Desimone and Duncan (1995), Kanwisher and Wojciulik (2000) and references therein. The experiment reported in O'Craven, Downing, and Kanwisher (1999) involves attention to overlapping objects more or less centrally located. It would be hard to imagine that a single hardwired and retinotopically organized feature can account for such phenomena. On the other hand the architecture proposed here does account for such phenomena, since the priming produced by the evoked model enhances the response to that class everywhere in the scene.

We note that feature based attention as observed in Treue and Martinez (1999) and McAdams and Maunsell (2000) can be viewed as a very simple form of object based detection. The 'model' contains only one feature.

6.1.1. Response time in target detection tasks

Object based attention can also be used to interpret various experimental findings regarding target detection among distractors (see Eriksen & St. James, 1986; Horowitz & Wolfe, 1998; Wolfe, 2000, 2001), where a range of response time slopes is observed. This is explained by the model above in terms of *false positive rates*. A similar explanation is suggested in Eriksen and St. James (1986). Depending on the distractor population there will be a certain probability P of false positives per distractor; The probability that an element from the distractor population will contain more than ρ of the feature location pairs of the target class model. This probability is hard to predict but is easy to measure empirically given a model and a sample of the distractor class (see for example experiments in Fig. 9). Note that the object model is quite coarse and may hit other object classes quite frequently. For example the $\mathscr{E}$ model, at the level of detail used in the reported experiment hits the population of $\mathscr{S}$'s with very high probability. However on the population of $\mathscr{N}$ distractors P is much lower.

The initial processing using the top-down priming will produce a number of locations above threshold. On average this will be PN, if N is the number of distractors and P the false positive probability. If these are serially processed using covert attention, and the visual system is able to avoid repetitions (i.e. revisiting a previously processed candidate location), this can be viewed as sampling *without replacement* from $PN + 1$ objects only one of which is the desired one. A straightforward computation shows that the expected search time would then be $(P \cdot N \cdot S)/2$, where S is the time required to shift covert attention to a location and process the data there through the W layer and the recognition layer A. If there

Fig. 10. Example of face detection and classification against non-face. The black dots denote all detections found in the image. The squares denote those detections that were classified as face.

is no memory for previously visited locations, as proposed in Horowitz and Wolfe (1998), then this corresponds to sampling *with replacement*. In this case there is no difference between a static image where candidate locations are repeatedly reproduced and one is chosen at random from among the candidates, and images that change every 20 ms. The expected search time in both cases would be $P \cdot N \cdot S$. In both search paradigms the slope depends on P and one can therefore expect a continuous range of slopes.

6.1.2. Efficient detection of unfamiliar shapes

The experiments reported in Wolfe (2001) have a very small response time slope for detecting a complex and relatively *unfamiliar* shape from among *familiar* distractors. For example detecting an upside-down elephant among upright ones, or an inverted 5 among regular 5's. In the current architecture this could be accommodated if the model evoked in W is that of the *distractor*, which is familiar—it has been learned. Instead of priming for the model in the replica module the opposite effect of suppression occurs. This will reduce activity in the U layers where data corresponding to the familiar distractor is present, leaving higher activity only in those layers where the data is different from the distractor. There will be many such layers, for example any layer corresponding to a blank window or only partially intersecting one of the objects. We can assume those are still rejected due to the low activity. In that respect there is probably very little dependence on the nature of the target itself. When the distractors are unfamiliar, this strategy will not work since their model is not yet learned, and the system would revert to using the target model.

6.2. Attentional modulation of neuronal responses

6.2.1. Lower levels: V4 and the F layers

In all retinotopic regions one finds neurons selective to the same feature with more or less the same receptive field stacked in columns. In terms of the copies of F layers used as input to the system (see Fig. 4), a column

corresponding to location y and feature f is simply $\{F^z_{f,y},\ z \in G\}$. One possibility is that these copies are used to implement location based and object based attention. The only difference between units in the same column is in their connections to and from higher areas in cortex, i.e. W and S. Such differences would not be observed by simple recordings of preferred stimuli.

Some evidence for such phenomena has been found in V4 (Connor et al., 1996). The authors report finding a large number of orientation selective neurons in V4 that change their response level to a stimulus in their classical receptive field, as a function of the location of attention. Attention was directed using other stimuli outside the receptive field of the neuron. In our model, if a stimulus of type f is shown at y and if the attended location x is selected in S, then the units $\{F^z_{f,x+z},\ z \in G\}$ (the diagonal slice in Fig. 4) are primed and all others are suppressed. In this case the *only* unit in the column $\{F^z_{f,y},\ z \in G\}$, expected to respond (the intersection of the two sets) would be $F^{z'}_{f,y}$, where $z' = y - x$. In other words, when a location for attention is selected, the response of units in a column of F would depend on the relative displacement z' between the selected location x and the receptive field location y, as observed in Connor et al. (1996).

Equating F with V4 is a possibility. We note that in our system the image array L is 100×100, with 40 features and 400 locations on the reference grid, yielding on the order of 10^8 units. This is definitely compatible with the number of units in area V4. Probably the dimensions of the array in V4 are smaller whereas there is a larger number of more complex features. It may however be possible that the same organization is present in V2 and even V1. In other words all three retinotopic areas compose the replica layer. The lower level areas providing higher resolution with simpler features and the higher level areas lower resolution with more complex features. After all it may be more economical to create models combining complex features at low spatial accuracy with simple features at higher accuracy.

Evidence for the priming required for location based attention is reported even in the absence of the expected stimulus in Luck, Chelazzi, Hillyard, and Desimone (1993) and Kastner, Pinsk, De Weerd, Desimone, and Ungerleider (1999). For single feature based attention, priming has been reported in MT (Treue & Martinez, 1999) and in V4 (McAdams & Maunsell, 2000).

6.2.2. Posterior IT and the W layer

The W layer could correspond to the posterior part of IT, also denoted TEO in Tanaka, Saito, Fukada, and Moriya (1991). This region exhibits neurons with responses to relatively simple features but with very large receptive fields. This is consistent with the properties of the W neurons. A unit (f,z) in W is summing the input from $F^z_{(f,x)}$ *for all* x (see Fig. 4), and would thus appear to have a receptive field as big as the visual field. This implies in fact that without a choice of attended location, the responses of these neurons convey very little information apart from the feature type. When a location is selected most of the input to the neuron is suppressed and only data in the selected location passes through. The receptive field appears to shrink around the selected location and the neurons seem to ignore their preferred stimulus outside this location. This phenomenon has been observed in IT cells in Sato (1988).

Indirect evidence for this can be found Chelazzi et al. (1993). Neurons selective for two different targets are found in IT. The monkey is presented with one target and required to detect it in a display with both targets present. After the presentation of the test display, activity of both neurons increases, however some 100–150 later the activity of the neuron selective for the wrong target decreases. In terms of our model the two neurons are in W, each one corresponding to a feature location pair found on one object and not on the other. Before a location is selected these neurons see all the activity in the replica module and hence are both activated upon presentation of the display. Once the detection process has located the correct target, that location is selected and the neuron selective for the wrong target no longer sees its preferred stimulus.

6.2.3. Is W also in V4?

Similar 'shrinking' of the receptive field has been observed in V4 cells in Moran and Desimone (1985). Furthermore the response of these cells is not affected when attention is directed outside their receptive fields. By contrast, in Motter (1993), McAdams and Maunsell (2000) and Connor et al. (1996) V4 cells are found that *do* modulate their response when the location of attention is outside their receptive fields. These findings could possibly be reconciled using a slight modification of the architecture described above. Instead of having W interact directly with the replica module, which admittedly requires a large number of long range connections, one could have intermediate W modules covering only parts of the field of view.

Assume for simplicity just two modules W^1 and W^2 both with the full array of units assigned to W in the architecture (see Fig. 11). The field of view L is divided in two L^1, L^2. Each $W^1_{f,z}$ unit of W^1 receives input and provides input to the $F^z_{f,x}$, $x \in L^1$, and each $W^2_{f,z}$ unit of W^2 is similarly connected to the L^2 section of F^z_f. Furthermore the L^1 units of S feed into all units of W^1 and the L^2 units of S feed into the units of W^2. The original unit $W_{f,z}$ in W provides input to both $W^1_{f,z}$ and $W^2_{f,z}$. Top-down model priming in W is directly mediated by W^1 and W^2. $W_{f,z}$ also receives input from these two units if no location has been selected. A simple screening mechanism is introduced so that if a location is selected in $L^1(L^2)$ only input from $W^1_{f,z}(W^2_{f,z})$ passes through to $W_{f,z}$. The screening can easily be implemented with an

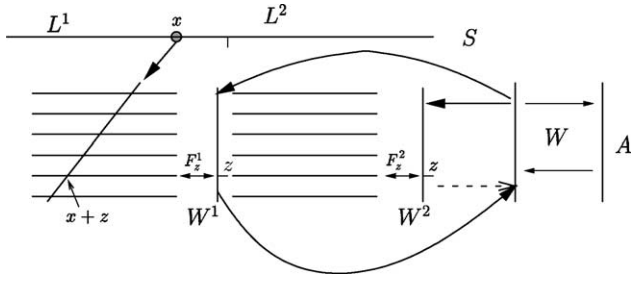

Fig. 11. Splitting W. Intermediate W^1, W^2 layers mediate location selection determined in S. A location is selected in L^1 (dark circle). Units in $W^1_{f,z}$, $W^2_{f,z}$ sum all activity in their corresponding part of F^z_f. A screening mechanism passes on only input from W^1 into W (bold arrow vs. dashed arrow). Top-down information for object based attention is passed from W to W^1 and W^2, which subsequently prime their respective parts of the replica module.

extra layer and we omit the details. This architecture is described in Fig. 11.

The receptive fields of W^1, W^2 are half the size of those of W (half the field of view). When a location is primed in L^1 it has no effect on the input into a unit in W^2, whereas the units in W^1 behave as if their receptive field has shrunk. Since the selected location is in L^1 the final W layer only receives the input from W^1. This is the data that ultimately gets processed through the connections between W and A.

The intermediate level could contain a larger number of W^i's with smaller and overlapping receptive fields. In terms of this formalism it may well be that these intermediate W^i layers are actually a subset of V4, corresponding to the type of neurons observed in Moran and Desimone (1985). By contrast other neurons in V4, with smaller receptive fields, may correspond to the simpler type of units $F^z_{f,x}$ that respond to the presence of a particular feature in a particular subregion of the field of view. These units would not show changes in receptive field size, rather they would show higher or lower activation depending on the attentional task as discussed above in Section 6.2.1.

6.2.4. Anterior IT and the A layer

The A layer contains neurons that would only respond to a more global arrangement of features, because they sum the input on a sample of neurons from W. Since the sample of neurons feeding into a unit of layer A is random, it is not easy to predict what are the particular shape characteristics it responds to. There is extensive literature on neurons in anterior IT responding to particular object classes, with significant invariance to translation, scale, occlusion, etc. Efforts have been made to identify the simplest shape such neurons would respond to (see Fujita, Tanaka, Ito, & Cheng, 1992; Ito, Tamura, Fujita, & Tanaka, 1995; Kobatake & Tanaka, 1994; Tanaka, 1996). However if indeed these neurons behave like the A layer neurons it would be difficult to precisely characterize their responses. Furthermore responses of these units may be modified by training.

For example in Sigala and Logothetis (2002) neurons in anterior IT are found that, after training, respond selectively to features that are discriminating between two categories of cartoon faces. They did not find neurons that respond selectively to uninformative features. A possible explanation is that these are indeed 'abstract' A type units. Each of the four features f_i, $i = 1, \ldots, 4$, in that experiment had three states. We represent this as 12 binary variables $f_{i,j}$, $i = 1, \ldots, 4$, $j = 1, 2, 3$, and assume they are represented in the W layer. Let $a \in A$ be a unit assigned to class 1, and assume a has connections to all 12 units. If say $f_{1,1}$ and $f_{1,2}$ are on with higher probability in class 1 than in class 2, then the synapse connecting them to a will have high efficacy, after training (see Section 4.2). If $f_{1,3}$ is of low probability its efficacy will remain more or less at the original state before training. When stimuli are presented to the network, those for which $f_{1,1} = 1$ or $f_{1,2} = 1$ will produce a higher response in a than those for which $f_{1,3} = 1$.

On the other hand if for example $f_{2,1}$, $f_{2,2}$ and $f_{2,3}$ (corresponding to the three states of feature 2) have more or less the same probability on the two different classes, the synapses connecting them to a will all have more or less the same efficacy, perhaps somewhat larger than the original efficacy before training. Thus the response of a to stimuli with different values of feature f_2 will be the same. In Section 6.5 we discuss some related predictions.

6.3. Invariant recognition without attention?

There is an interesting experiment reported in the literature pointing to the possibility that recognition may occur without attention (Li, VanRullen, Koch, & Perona, 2002). They describe rapid classification between animal and non-animal photographs, presented in the periphery, while performing a task requiring high attentional load at the fixation point. (A related experiment with two simultaneous displays is described in Rousselet, Fabre-Thorpe, and Thorpe (2002).) Subjects perform well on this task after extensive training, but are unable to perform other peripheral tasks such as distinguishing between a T and an L. The same phenomenon is observed for vehicle non-vehicle classification. The conclusion drawn by the authors of this paper is that the system is able to perform a very high level task very quickly and without attention. This indeed seems to put in question much of what we have discussed here, as well as the basic assumption repeated in the literature on attention: there are limited resources and the visual system must use attention to direct these resources in an effective manner.

If the distinction between the two classes in this experiment is done at the 'object level' then a very complex

analysis is required, taking into account the multiple possible shapes with which animals (vehicles) can appear. This would seem then *a much higher level task* than discriminating between T and L. On the whole it is difficult to interpret these experiments because there is no systematic control over the population of photographs employed, and it is unclear what characterizes the errors in the discrimination tasks.

One explanation may have to do with the ability of the system to deal with two loci of attention. Some experiments seem to indicate such a possibility (Krammer & Hahn, 1995). This may also explain the results in Rousselet et al. (2002). Alternatively, the system may be able to quickly shift the covert attention between different locations. This may be facilitated by the fact that the peripheral photograph is presented some 60 ms after the discrimination task at the fixation point. (Recall also the experiment in Horowitz and Wolfe (1998) where displays were randomly shuffled every 110 ms with no major effect on performance.) The data then enters the ventral stream and gets processed sequentially. After all the response time in these experiments is quite long—1000 ms.

An alternative explanation is that the visual system is actually performing a very low level type of discrimination, such as center/surround for example. Typically the animals or vehicles will be centered in the photograph and will be in the foreground relative to the rest of the image. This interpretation is suggested in Li et al. (2002), and is tested by running a second experiment where half the non-animal images are photos of a vehicle. If performance does not decrease this would be an indication that the center/surround explanation should be rejected. It is difficult to judge the outcome of this second experiment. The data shown is not in terms of the actual percentage of correct answers, but rather using a rescaling relative to the rates without attending to a task at the fixation point. It is unclear if performance has remained the same in the second experiment. If the visual system still uses the center/surround strategy (in an ideal situation) this would still yield about 75% correct: the 50% animal displays, the 25% generic background displays. This is far higher than the random choice 50% outcome.

6.4. Detection and recognition in image sequences

A large body of literature describes RSVP experiments involving various detection and recognition tasks in a sequence of rapid presentations of isolated objects, see for example Potter and Chun (1995) and Biederman (1995). Here the complicating factor is not location, the objects are presented alone in the image. In our model, since the data in the replica layer U is always summed up into W (see Figs. 1 and 4), even without location selection, there is no place for confusion or ambiguity if only one object is present in the scene. The data ends up in W and is classified in A. The challenge is to explain how the system deals with the short time intervals between presentations (on the order of 100 ms), how the sequence of images is channeled through the processing stages without interference, and how attention affects performance. These issues would require the introduction of more realistic time dynamics into the model and are beyond the scope of this paper.

6.5. Predictions

6.5.1. Attention

The model described here leads to several rather straightforward predictions. First in the context of location based attention, essentially as an extension of the experiment in Connor et al. (1996), we expect units in a single vertical column in area V4 and perhaps V2 and V1, corresponding to the same location and feature, to have different responses to the preferred stimulus in their common receptive field. Specifically, each neuron in the column is expected to have a *preferred* locus of attention. When attention is directed to that location the unit exhibits the strongest response to the preferred feature present in its receptive field. This locus of attention would probably change gradually over certain intervals of a vertical probe into the column.

Second, in the context of object based attention, when the task is to find a target in a cluttered scene, the model employs top-down priming of units in F from W (see Figs. 2 and 4). This implies that in the presence of the preferred stimulus different units in a column would exhibit different levels of activity, depending on whether the object model has the preferred feature at their preferred object centered coordinate, i.e. the z index used in Section 2.3. Alternatively the response of a given unit to its preferred stimulus should change as the target object is changed.

Furthermore, when object based attention is in effect, we expect increased activity among neurons throughout retinotopic layers such as V1, V2 or V4, even in the absence of the stimulus. This activity corresponds to the priming of the object model at all possible shifts.

6.5.2. Recognition

The discussion in Section 6.2.4 hypothesized that the neurons observed in anterior IT correspond to the abstract classification layer A. This implies that responses of these neurons are shaped by training and can be changed. For example in the experiment reported in Sigala and Logothetis (2002) one could imagine creating a new partition into two classes, where the old discriminating features are no longer informative and new ones are. We then expect that after training, the same neuron that was selectively responsive to different values of one feature would become selectively responsive to a

new more informative feature. Such an experiment would be of particular interest, since it would point to the possibility that the preference of neurons in anterior IT to particular objects or features is not a permanent attribute, but rather one that can change with training.

7. Conclusion

We have presented an architecture for translation invariant object detection and recognition using a replica module containing multiple copies of retinotopic feature arrays properly wired to two higher level layers: a location selection layer S, and a model layer W. Priming either from S or from W are the mechanisms whereby the appropriate data is selected to be passed on. Learning is restricted to the synapses between W and an 'abstract' layer A that codes for the different classes in terms of random subpopulations. This model can be used to explain some psychophysical experiments and is consistent with attentional modulated responses in V4 and IT neurons reported in the literature. Some relations have been discussed to anterior IT neurons, which exhibit trained class selectivity.

We hypothesize that the columnar organization in visual cortex, where multiple units responding to the same feature at the same location are arranged in a column, could be precisely the copying mechanism needed for the proposed implementation of top-down object based and location based attention. This leads to simple predictions on varied responses within such a column depending on the selected location or the target object. Furthermore we hypothesize that neurons in anterior IT may change their selectivity as a function of training, and as such do not have a particular hard wired preferred stimulus.

The network described here is synthetic, the neurons are simplified binary on–off units and there is no real dynamics. Priming and competition are not obtained through more realistic dynamic mechanisms. Introducing dynamic interactions between the two learning processes described, between the detection and recognition processes, and between bottom-up and top-down location selection could very well give rise to interesting phenomena.

Acknowledgement

This work was partially supported by NSF ITR/0219016a.

References

Amit, Y. (2000). A neural network architecture for visual selection. *Neural Computation, 12*, 1059–1082.

Amit, Y. (2002). 2d Object detection and recognition: Models, algorithms and networks. Cambridge, MA: MIT Press.

Amit, Y., & Geman, D. (1999). A computational model for visual selection. *Neural Computation, 11*, 1691–1715.

Amit, Y., & Mascaro, M. (2001). Attractor networks for shape recognition. *Neural Computation, 13*, 1415–1442.

Biederman, I. (1995). Visual object recognition. In S. M. Kosslyn, & D. N. Osherson (Eds.), *Visual cognition* (pp. 121–166). Cambridge, MA: MIT Press.

Cheal, M., & Marcus, G. (1997). Evidence of limited capacity and noise reduction with single-element displays in the location-cuing paradigm. *Journal of Experimental Psychology: Human Perception & Performance, 23*, 51–71.

Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature, 363*, 345–347.

Chum, M. M., & Wolfe, J. M. (2000). Visual attention. In E. P. Goldstein (Ed.), *Blackwell Handbook of Perception*. Blackwell.

Connor, C., Gallant, J. L., Preddie, D. C., & Van Essen, D. C. (1996). Responses in area v4 depend on the spatial relationship between stimulus and attention. *Journal of Neurophysiology, 75*, 1306–1308.

Deco, G. A. (2000). A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition. *Vision Research, 40*, 2845–2859.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review, 87*, 272–300.

Eriksen, C. W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: a zoom lens model. *Perception and Psychophysics, 40*, 225–240.

Fleuret, F., & Geman, D. (2001). Coarse-to-fine face detection. *International Journal of Computer Vision, 41*, 85–107.

Fujita, I., Tanaka, K., Ito, M., & Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature, 360*, 343–346.

Fukushima, K., & Miyake, S. (1982). Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition, 15*, 455–469.

Fukushima, K., & Wake, N. (1991). Handwritten alphanumeric character recognition by the neocognitron. *IEEE Transactions on Neural Networks, 2*, 355–365.

Hedgé, J., & Van Essen, D. C. (2000). Selectivity for complex shapes in primate visual area v2. *Journal of Neuroscience, 20*, RC61.

Henderson, J. M. (1991). Stimulus discrimination following covert attentional orienting to an exogenuos cue. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 91–106.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature, 394*, 575–577.

Humphreys, G. W., & Heinke, D. (1998). Spatial representation and selection in the brain: neuropsychological and computational constraints. *Visual Cognition, 5*, 9–47.

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

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on PAMI, 20*, 1254–1260.

Kanwisher, N., & Wojciulik, E. (2000). Visual attention: insights from brain imaging. *Nature Reviews, Neuroscience, 1*, 91–100.

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention inn the absence of visual stimulation. *Neuron, 22*, 751–761.

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neuroscience, 71*(3), 856–867.

Krammer, A. F., & Hahn, S. (1995). Splitting the beam: distribution of attention over noncontiguous regions of the visual field. *Psychological Science, 6*, 381–386.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *PNAS, 99*, 9596–9601.

Luck, S., Chelazzi, L., Hillyard, S., & Desimone, R. (1993). Effects of spatial attention on responses of v4 neurons in macaque. *Society of Neuroscience Abstract, 19*, 27.

McAdams, C. J., & Maunsell, J. H. R. (2000). Attention to both space and feature modulates neuronal responses in macaque area v4. *Journal of Neurophysiology, 83*, 1751–1755.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science, 229*, 782–784.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas v1, v2 and v4 in the presence of competing stimuli. *Journal of Neurophysiology, 70*, 909–919.

O'Craven, K., Downing, P., & Kanwisher, N. (1999). fmri evidence for objects as the units of attentional selection. *Nature, 401*, 584–587.

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

Potter, M. C., & Chun, M. M. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology, 21*, 109–127.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*, 1019–1025.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience, 5*, 629–630.

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

Sato, T. (1988). Effects of attention and stimulus interaction on visual responses of inferior temporal neurons. *Journal of Neurophysiology, 60*, 344–364.

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature, 415*, 318–320.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19*, 109–139.

Tanaka, K., Saito, H. A., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects and the inferotemporal cortex of the macaque monkey. *Journal of Neuroscience, 66*(1), 170–189.

Treisman, A., & Sharon, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 459–478.

Treue, S., & Martinez, J. C. (1999). Feature based attention influences motion processing gain in macaque visual cortex. *Nature, 399*, 575–579.

Viola, P., & Jones, M. J. (2002). Robust real time object detection. *International Journal of Computational Vision*.

Wilkinson, F., Wilson, H. R., & Habak, C. (1998). Detection and recognition of radial frequency patterns. *Vision Research, 38*, 3555–3568.

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

Wolfe, J. M. (2000). The deployment of visual attention. In Search and target acquisition. Vol. RTO. NATO-RTO, Utrecht, Netherlands.

Wolfe, J. M. (2001). Assymetries in visual search. *Perception and Psychophysics, 63*, 381–389.