A Computational Model for Visual Selection

Yali Amit * and Donald Geman †

February 1998

*Department of Statistics, University of Chicago, Chicago, IL, 60637; Email: amit@galton.uchicago.edu. Supported in part by the Army Research Office under grant DAAH04-96-1-0061 and MURI grant DAAH04-96-1-0445,

†Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003; Email:geman@math.umass.edu. Supported in part by the NSF under grant DMS-9217655, ONR under contract N00014-97-1-0249, Army Research Office under MURI grant DAAH04-96-1-0445.

Abstract

We propose a computational model for detecting and localizing instances from an object class in static grey level images. We divide detection into *visual selection* and *final classification*, concentrating on the former: Drastically reducing the number of candidate regions which require further, usually more intensive, processing, but with a minimum of computation and missed detections. Bottom-up processing is based on local groupings of edge fragments constrained by loose geometrical relationships. They have no *a priori* semantic or geometric interpretation. The role of training is to select special groupings which are moderately likely at certain places on the object but rare in the background. We show that the statistics in both populations are stable. The candidate regions are those which contain *global* arrangements of several local groupings. Whereas our model was not conceived to explain brain functions, it does cohere with evidence about the functions of neurons in V1 and V2, such as responses to coarse or incomplete patterns (e.g., "illusory contours") and to scale and translation invariance in IT. Finally, the algorithm is applied to face and symbol detection.

1 Introduction

Approximately 150 milliseconds after visual input is presented, or within several tens of milliseconds after local processing in V1, cells in IT signal that an object has been detected and a location has been selected in a field of view larger than the fovea. Assuming a specific detection task is required, the decision is rapid but might be wrong. Additional processing might reveal that the desired object is not in the vicinity of the first location and a sequence of locations may need to be inspected. Therefore, in a very short period of time, local information is processed in a region somewhat larger than the fovea in order to identify "hot spots" which are likely, though not certain, to contain a desired object or class of objects. Final determination of whether

these candidate locations correspond to objects of interest requires intensive high resolution processing after foveation. This scenario - *visual selection* (or *selective attention*) and sequential processing - is widely accepted in the literature; see (Thorpe, Fize & Marlot 1996, Desimone, Miller, Chelazzi & Lueschow 1995, Lueschow, Miller & Desimone 1994, Van Essen & Deyoe 1995, Ullman 1996).

In artificial vision, the problem of detecting and localizing all instances from a generic object class, such as faces or cars, is referred to as *object detection*. Our goal is an efficient algorithm for object detection in static grey level scenes, emphasizing the role of visual selection. By this we mean quickly identifying a relatively small set of poses (position, scale, etc.) which account for nearly all instances of the object class in an image. Experiments are presented illustrating visual selection in complex scenes, as well as the final classification of each candidate as "object" or "background." We also explore connections between our computational model and evidence for neuronal responses to "illusory" contours or otherwise incomplete image structures in which fragmentary data is sufficient for activation. We argue that, due to spatial regularity, it is more efficient and robust not to fill in missing fragments.

Here is a synopsis of the approach: Bottom-up processing is based on local features defined as flexible groupings of nearby edge fragments. The object class is represented by a union of global spatial arrangements, this time among several of the local features and at the scale of the objects. Photometric (i.e., grayscale) invariance is built into the definition of an edge fragment. Geometric invariance results from explicit disjunction (ORing): The local groupings are disjunctions of conjunctions of nearby edge fragments and the global arrangements are disjunctions of conjunctions of the local ones. In principle we entertain *all* possible local features, a virtually infinite family. The role of training is to select dedicated local groupings which are each rare in the "background population" but moderately likely to appear in certain places on the object. We will provide evidence that a very small amount of training data may suffice to identify such groupings.

Visual selection is based on an image-wide search for each global arrangement in the union over a range of scales and other deformations of a reference arrangement. Each instance signals a candidate pose. Accurate visual selection is then feasible due to the favorable marginal statistics and to weak dependence among spatially distant groupings. It is fast because the search is coarse-to-fine and the indexing in pose space is driven by rare events, namely the global arrangements; in addition, there is no search for "parts" (or other sub-classification task) and no segmentation per se. The result of an experiment in face detection is shown in Figure 1. The lefthand panel shows the regions containing final detections. The righthand panel is a grayscale rendering of the logarithm of the number of times each pixel in the image is accessed for some form of calculation during visual selection; the corresponding image for many other approaches, e.g., those based on artificial neural networks, would be constant.

Part of this program is familiar. The emphasis on groupings and spatial relationships, the use of edges to achieve illumination invariance, the general manner of indexing and the utility of statistical modeling have all been explored in object recognition; some points of contact will be mentioned shortly. Moreover, the general strategy for visual selection goes back at least to (Lowe 1985) and others who emphasized the role of selecting groupings based on their statistical or "non-accidental" properties.

What seems to be new is that our approach is *purely* algorithmic and statistical. The groupings have no a priori semantical or geometrical content. They are chosen within a very large family based solely on their statistical properties in the object and background populations. They are also more primitive and *less* individually informative than the model-based features generally found in computer vision algorithms. For example, we use the term "edge fragment" even though the marked transitions have no precise orientation. Moreover, the groupings do not necessarily correspond to smooth object contours and other regular structures (such as corners and lines) that are often the target of bottom-up processing. In other words, there is no geometrical

or topological analysis of contours and object boundaries. (See Figure 3.) Nor is there an abstract concept of a "good grouping" as in Gestalt psychology.

In addition, we argue that visual selection, if not final classification, can be accomplished with object representations which are very coarse and sparse compared with most others, for example 3D geometric models, structural descriptions based on "parts" (Winston 1970, Biederman 1985) and "pictorial representations" (Ullman 1996). The "face graphs" in (Maurer & von der Malsburg 1996) are closer in spirit, although the "jets" (outputs from multiple Gabor filters) at the graph vertices are more discriminating than our local groupings; also, the representation there is much denser, perhaps because the application, namely face *recognition*, is more challenging.

Our representation of pose space (a three point "basis" or local coordinate system) is the same as in geometric hashing (Lamdan, Schwartz & Wolfson 1988), wherein the local features are affine invariants (e.g., sharp inflections and concavities) and objects are represented by hash tables indexed by feature locations. But again our framework is inherently nondeterministic: Features may or may not be visible on the objects, regardless of occlusion or other degrading factors, and are characterized by probability distributions. In addition, the global arrangements are more than a list; it is the geometrical constraints which render them "rare" in the background population. The statistical framework in (Rojer & Schwartz 1992) is similar, although they do not suggest a systematic exploration of local features. Finally, there are shared properties with artificial neural networks (Rowley, Baluja & Takeo 1998, Sung & Poggio 1998), for example the emphasis on learning and the absence of formal models. However, our algorithm is not purely "bottom-up" and our treatment of invariance is explicit; we do no expect the system to learn about it, or about weak dependence or coarse-to-fine processing. These properties are "hard-wired."

In the following section the object detection and visual selection problems are formulated more carefully. In Section 3 we delineate the statistical and invariance properties we require of our local and global features. The local edge groupings and

Figure 1: Left: Regions containing final detections. Right: A grayscale rendering of the *logarithm* of the number of times each pixel in the image is accessed for some form of calculation during visual selection.

global arrangements are defined in Section 4. Training and object representations are discussed in Section 5. In Section 6 we describe how to search for these representations and identify candidate regions in an invariant manner; final classification of these regions as object or background is explained in Section 7. Section 8 is devoted to a statistical analysis of the features, especially their densities in natural images, which motivates the choice of particular parameter values and allows us to estimate error rates. In Section 9 we present some experiments on face and symbol detection, demonstrating some robustness to occulsion during the selection stage. Section 10 is devoted to connections with brain modeling, especially evidence for similar types of coarse processing in the visual cortex and the role of grouping and segmentation; we also comment briefly on suitable neural network-type architectures for efficient implementation. The final section summarizes the main strengths and weaknesses of the proposed model.

2 Problem Formulation

The problem is to detect objects of a specific class, e.g., faces, cars, a handwritten "5", any digit, etc. In order to narrow the scope we assume static gray level images, and hence do not utilize color, depth or motion cues. However, since our initial processing is edge-based, one way to incorporate such information would be to replace intensity edges by those resulting from discontinuities in color, depth or motion. Moreover, we do not use *context*. Thus, the detection is primarily shape-based.

We assume that the object appears at a limited range of scales, say $\pm 25\%$ of some mean scale, and at a limited range of rotations about a reference orientation (e.g., an upright face). Other poses are accommodated by applying the algorithm to pre-processed data; for example we detect faces at scales larger than the reference one by simple downsampling.

We want to be more precise about the manner in which a detected object is localized within the image. Since the given range of scales is still rather wide and since we also desire invariance to other transformations, for instance local linear and nonlinear image deformations, it is hardly meaningful to identify the pose of an object with a single degree of freedom. Instead we assign each detection a *basis* - three points (six degrees of freedom) which define a local coordinate system. Consequently, in addition to translation, there is an adjustment for scale and other small deformations. Of course this extended notion of localization increases the number of poses by several orders of magnitude; within the class of transformations mentioned above, the number of bases in a 100×100 image is on the order of ten million.

Assume that each image in a training set of examples of the object is registered to a fixed reference grid in such a way that three distinguished points on the object are always at the same fixed coordinates, denoted z_1, z_2, z_3. As an example of three distinguished points on a face, consider the "centers" of the two eyes and the mouth. Typically we use a reference grid of about 30×30 pixels and expect the smallest detection to be at a scale of around 25×25. Each possible image basis (b_1, b_2, b_3)

then determines a unique affine map which carries z_i to b_i for $i = 1, 2, 3$. In addition, the reference grid itself is carried to a subimage, or "region-of-interest" (ROI), around the basis.

The ROI plays the role of a segmented region. In particular, there is no effort to determine a silhouette or a subregion consisting more or less exactly of object pixels. Note also that we do *not* search directly for the distinguished points; they merely define localization. We find that a search for either a silhouette or for special points during a chain of processing leading up to recognition is highly unreliable; in fact, it may only be when the object as a whole is detected that such attributes can actually be identified.

Visual selection means identifying a set of candidate ROIs; the ultimate problem is to classify each one as "object" or "background," which may not be easy with high accuracy. However, given the drastic reduction of candidates, presumably the final classification of each candidate could be allotted considerable computational resources. Moreover, this final classification can be greatly facilitated by *registering* the image data in the ROI to the reference grid using the affine map mentioned above. For example, in our previous work, the final classification was based on training decision trees using registered and normalized gray level values, and the computer vision literature is replete with other methods, such as those based on neural networks. However, this is not the main focus of this paper. *The theme here is the reduction of the number of ROI's which require further and intensive processing from several millions to several tens, and with a minimum of computation and missed detections.*

3 Feature Attributes

Our local features are binary, point-based image functionals which are defined modulo translation. Moreover, the set of all occurrences on an image-wide basis is regarded as the realization of a *point process*, assumed to be stationary in the background

population in a statistical sense. Instances of this process have no a priori semantic interpretation and hence there is no sub-recognition problem implicit in their computation. *In particular there is no such thing as a "missed detection" at the feature level.* Their utility for visual selection depends on the following attributes:

- **LI: Stability:** A significant degree of invariance to geometric deformations and to gray level transformations representing changes in illumination.
- **LII: Localization:** Appearance in a specified small region on a significant fraction (e.g., one-half) of the registered training images of the object.
- **LIII: Low Background Density:** Realizations of the point process should be relatively sparse in generic background images.

The first two properties are linked. Suppose, for example, that all images of the object corresponded to smooth deformations of a template. Then stability would imply that a local feature which was well-localized on the template should be present near that characteristic location on a sizable fraction of the examples. In the next section we will exhibit an enormous family of local features with property **LI**, in Section 5 we will explain how to select a small subset of these based on training data which satisfy **LII**, and in Section 8 we show to select the model parameters in order to achieve **LIII**.

Global information is essential. Complex objects are difficult to detect (and distinguish from one another) even when coherent "parts" are individually recognized, and recognizing parts independently of the whole object is itself a daunting challenge. For example, although faces can be detected at low resolution, it might be very difficult to identify say a left eye based only on the intensity data in its immediate vicinity, i.e., outside the context of the entire face; see the example and discussion in (Ullman 1996). Furthermore, local features do not provide information about the pose, except for translation.

A *global arrangement* in a registered training image is the conjunction (simultaneous occurrence) of a small number of local features subject to the constraint that their locations in the reference grid are confined to specified regions. An instance of a global arrangement in a test image occurs in the ROI of a basis if the locations of the local features fall in their distinguished regions *in the local coordinate system determined by the basis*. This will be made more precise later on. The properties we need are these:

- **GI: Coverage:** A small collection (union) of such arrangements "covers" the object class in the range of scales and rotations in which the object is expected to appear in the scene.

- **GII: Rare Events:** The arrangements are very rare events in a generic scene, i.e., in general background images.

The precise meaning of **GI** is that a very high percentage of images of the object exhibit *at least one* global arrangement after registration to the reference grid. In other words, the union of the arrangements is nearly an invariant for the object class. During selection, the object instances which are detected are those which are covered by at least one global arrangement. Hence this "coverage probability" is lower bound on the false negative rate of the entire detection process. The coverage probability is directly determined by the joint statistics of the local features on registered images of the object class, together with the degree of invariance introduced in the definition of the arrangements, i.e., the amount of "slack" in the relative coordinates of the local features; see Section 8.

Property **GII** - limiting the number of "hot spots" - is of course related to false positive error, as will be explained more fully in Section 8. Statistical characteristics of the global arrangements in "natural scenes" are determined by the density and higher order moments of the point processes corresponding to the local features.

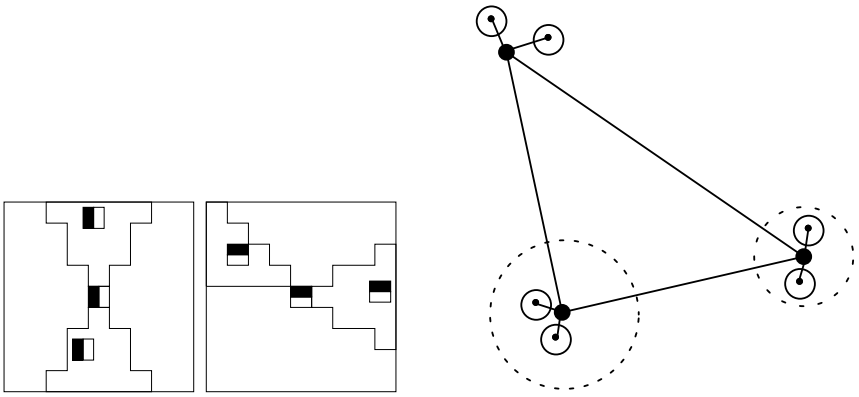

Figure 2: Left: Two examples of local edge groupings with $N_{edges} = 2$ edges in addition to the center one, each allowed to lie anywhere in a subregion of size $N_{pixels} \approx 10$. Right: A global grouping of three local ones; the small circles represent the subregions in the edge groupings and the large dotted circles represent the analogous subregions for the global arrangement; see 4.2).

4 Groupings

All features presented below are defined in terms of coarsely oriented edge detectors. A great many edge detectors have been proposed and some of these with enough grayscale invariance would suffice for our purposes. The one we use is based on comparisons of intensity differences and is consequently invariant to linear transformations of the grey scale, insuring the photometric part of **LI**. There are four edge types, corresponding roughly to vertical and horizontal orientation and two polarities; the details are in (Amit, Geman & Jedynak 1998) and are not important for the discussion here, except to note that the orientation is not very precise. For example, the "vertical" edge responds to any linear boundary over a ninety degree range of orientations.

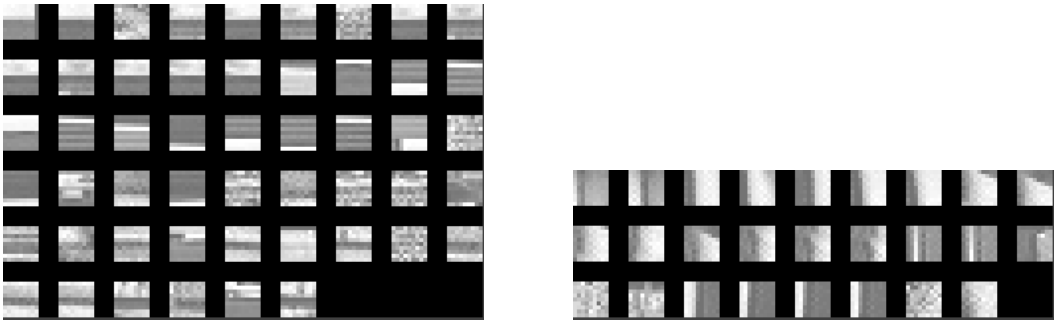

Figure 3: Examples of 9×9 subimages centered at instances of local features (edge groupings) identified for faces. Left: Samples of one local feature from an image without faces. Right: The same thing for another local feature.

4.1 Edge Groupings

The local features are flexible spatial arrangements of several edge fragments, organized as disjunctions of local conjunctions of edges. Each feature is defined in terms of a "central edge" of some type, and a number N_{edges} of other edge types which are constrained to lie in specific subregions within a square neighborhood of the location of the center edge. The local feature inherits the location of the central edge. The sizes of the subregions are all the same and denoted by N_{pixels}. Typically the subregions are wedge-shaped as indicated in Figure 2. Disjunction - allowing the N_{edges} edges to float in their respective subregions - is how geometric invariance (**LI**) is explicitly introduced at this level; there is also disjunction at the global level as indicated earlier.

The frequency of occurrence of these groupings depends on N_{edges}, N_{pixels} and the particular spatial arrangement. Among the set of all possible edge groupings - the generic feature class - most are simultaneously rare in both object and background images. When specific groupings are selected according to their frequency in training examples of a particular object, they appear to be loosely correlated with evidence for contour segments, or even relationships among several segments. In Figure 3 we show subimages of size 9×9 which contain two particular groupings common

Figure 4: Left: All instances of horizontal edges. Right: All instances of a local feature dedicated to faces.

in faces. The one on the left is typically located at the region of the eyebrows; the grouping involves some horizontal edges of one polarity above some others of the opposite polarity. These instances were chosen randomly from among all instances in a complex scene with no faces.

The point process determined by any local feature, as localized by the central edge, is a *thinning* of the point process determined by instances of the central edge type. Each additional edge type in the grouping, and corresponding subregion thins it even further. Figure 4 illustrates the thinning by showing all instances of horizontal edges of one polarity alongside all instances of a local feature centered at the horizontal edge with $N_{edges} = 3$ and $N_{pixels} = 10$.

4.2 Global Groupings - Triangles

Global groupings are defined in a similar manner to the local groupings. The edges are replaced by entire local groupings, and the distances between the features can vary in a much larger range. The degree of geometric invariance is again determined by the degree of disjunction, which in turn depends on the size of the subregions in which the local groupings are constrained to lie.

We will concentrate on global arrangements of exactly three local features, referred to as "triangles." (This is the minimum number necessary to uniquely determine a

basis.) Let us be more specific about what it means for a particular triangle Δ - triple of local features - to be present "at pixel x_0". Denote the "central" local feature by α_0 and the two others by α_1 and α_2. Of course α_0, α_1 and α_2 are each local groupings of edges. Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be two boxes centered at the origin; these determine the degree of disjunction for α_1 and α_2. Also, let v_1 and v_2 be two vectors; these determine the locations of the boxes relative the location of α_0, in other words, the overall shape of the arrangement. Then there is an instance of the triangle Δ at x_0 if feature α_0 is present at x_0, feature α_1 is present at some point $x_1 \in x_0 + v_1 + \mathcal{B}_1$ and feature α_2 is present at some point $x_2 \in x_0 + R_{x_1 - x_0} v_2 + \mathcal{B}_2$, where $R_{x_1 - x_0}$ is the rotation determined by the vector $x_1 - x_0$. The size of $\mathcal{B}_1$ is set to accommodate the range of scales at which the triangle can occur. Once the second point of the triangle is found, the scale is determined and $\mathcal{B}_2$ accommodates the residual variability. (See Figure 2.)

5 Object Representations and Training

Let $\mathcal{L}$ denote the training set of images. We first compute the edge (fragment) map for each member of $\mathcal{L}$ and then register these maps to a fixed size reference grid, as described in Section 2. In this way, linear variability is essentially factored out. We are going to induce a collection $\alpha_i, i = 1, \ldots, N_{types}$, of local edge groupings, each of which is "common" in a certain region of the reference grid (equivalently, of the object). Recall that N_{edges} denotes the number of edges in the grouping in addition to the central edge and N_{pixels} denotes the size of the regions in which the edges are allowed to "float" (see Figure 2). Fix N_{edges}, N_{pixels} and let $\mathcal{R}$ be a set of candidate regions - small, wedge-shaped neighborhoods of the origin.

1. Set feature counter $I = 0$. Loop over disjoint 5×5 boxes on the reference grid. For each box B:

   (a) For each possible combination (e_0, e_1, R_1), where e_0, e_1 are any possible edge types and $R_1 \in \mathcal{R}$, count the number of training points in $\mathcal{L}$ for

which an instance of the triple occurs in B. This means e_0, the central edge, is located anywhere in B and e_1 is located anywhere in R_1 relative to the location of e_0. Pick the triple with highest count and let $\mathcal{L}_1$ denote the set of data points which have an instance of this triple in B. For each data point $d \in \mathcal{L}_1$, let $x_{d,t}, t = 1, \ldots, n_{d,1}$ denote all locations of the first edge e_0 for which the chosen triple was found. Set $j = 2$.

(b) Loop over all possible pairs e_j, R_j and count how many data points $d \in \mathcal{L}_{j-1}$ have an edge of type e_j anywhere in the subregion R_j relative to one of the locations $x_{d,t},\ t = 1, \ldots, n_{d,j-1}$. Find the pair with highest count and let $\mathcal{L}_j \subset \mathcal{L}_{j-1}$ denote the data points which have an instance of this pair. For each $d \in \mathcal{L}_j$, let $x_{d,t}, t = 1, \ldots, n_{d,j}$ denote all the locations of the first edge for which the pair was found.

(c) $j \leftarrow j + 1$. If $j < N_{edges}$ goto (b).

2. If $|\mathcal{L}_{N_{edges}}|/|\mathcal{L}| > \tau$, record the feature $\alpha_I = (e_0, e_1, R_1, \ldots, e_{N_{edges}}, R_{N_{edges}})$ at the center of B, say y_I. All data points in $\mathcal{L}_{N_{edges}}$ have an instance of e_0 at a location $x \in B$ and an instance of e_i in region R_i relative to x for each $i = 1, \ldots, N_{edges}$.

   Set $I \leftarrow I + 1$.

3. Move to the next box and goto 1.

We end up with I local features α_i at locations y_i. Typically I will be larger than N_{types} and we choose a subset of size N_{types} for which the locations y_i are "spread out" over the object. By requiring τ to be sufficiently large (e.g., $\tau = .5$), we establish the "localization property" **LII**. This is the *only* training which takes place for visual selection. The time required is on the order of *minutes* for several hundred training images.

Each triple $(i, j, k), 1 \leq i < j < k \leq N_{types}$, of selected local features determines a "model" triangle $\Delta = \Delta_{ijk} = (y_i, y_j, y_k)$. *The set of these triangles is the object repre-*

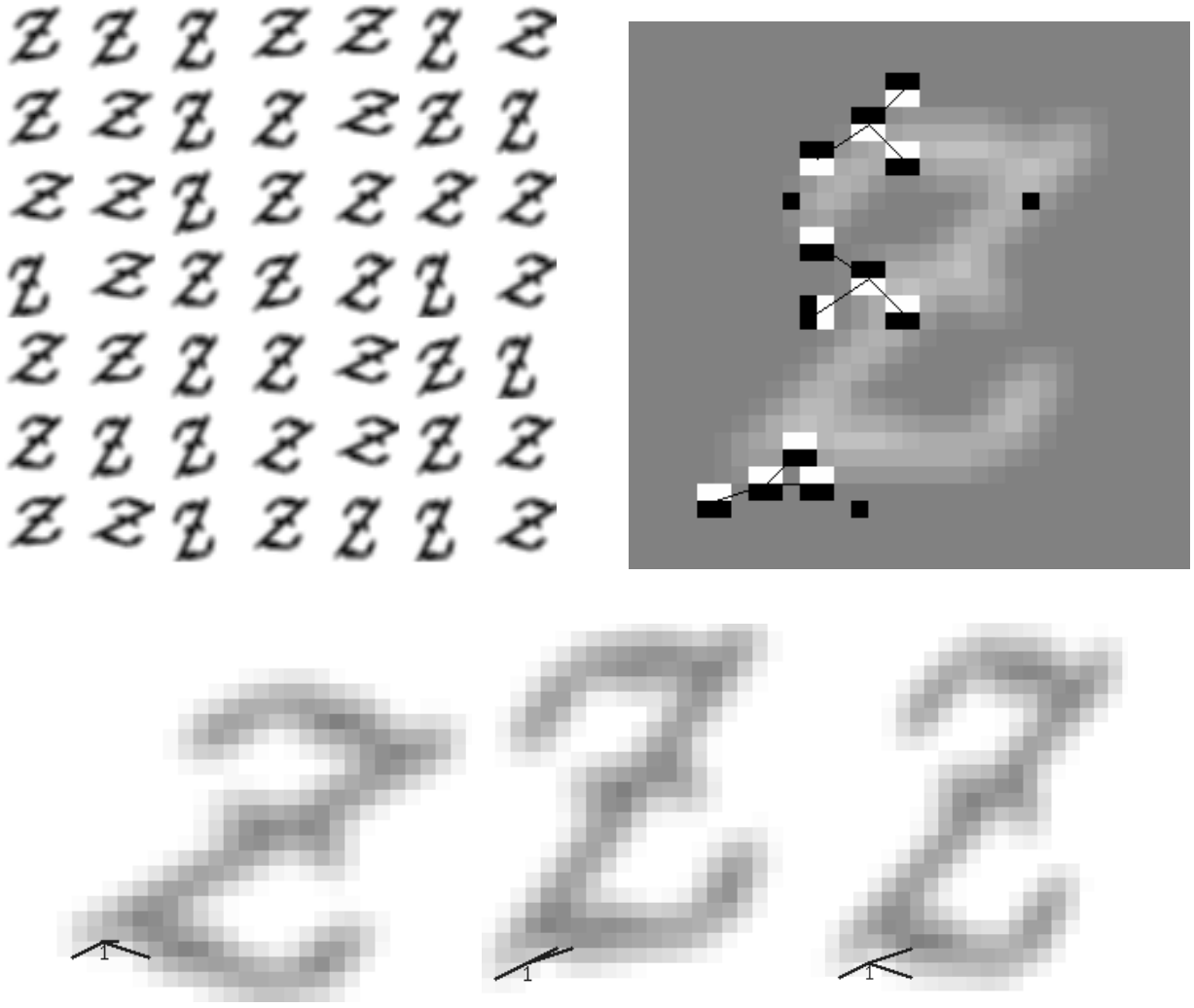

Figure 5: Top Left: A collection of randomly deformed $\mathcal{Z}$'s. Top Right: Three local features in their reference grid locations, superimposed on an image of the prototype $\mathcal{Z}$. The pairs of black/white rectangles denote an edge. Bottom: One instance of the bottom left local feature on three unregistered random $\mathcal{Z}$'s. They are all found at the correct location. Note the variability in the instantiation of the local feature.

sentation. In Figure 5 we show a collection of randomly deformed $\mathcal{Z}$'s, obtained from a prototype by applying a random low frequency non-linear deformation and then a random rotation and skew. We also show a smoothed version of the prototype (which is not part of the training set) in the reference grid. The three black dots indicate the basis points z_1, z_2, z_3 (see Section 2). Also superimposed are three local features identified for this class of objects at their model locations in the reference grid. Each pair of black/white rectangles denotes an edge at one of the four orientations. The three local features represent one of the triangles in the model. Note that the actual instances on training data vary considerably in their locations. However the invariance incorporated in the search for these triangles accommodates these variations. The bottom row shows three $\mathcal{Z}$'s with an instance of one of the features. The images are *not* registered and the feature was detected on the unregistered images. In a test data set of 100 perturbed symbols all of these local features were found in over 50% of the symbols at the correct location.

6 Invariant Search

The triangles provide a straightforward mechanism for incorporating invariance into the search for candidate bases. Given an image and a model triangle $\Delta = (y_i, y_j, y_k)$ for three local features $\alpha_i, \alpha_j, \alpha_k$, we search for all instances of these local features which form a triangle *similar* to the model triangle Δ up to small perturbations and a scaling of +/- 25%. The image-wide search for similar triangles is equivalent to a search for a global arrangement, see section 4.2 with $v_1 = (y_j - y_i), v_2 = (y_k - y_i)$, and the size of $\mathcal{B}_1$ and $\mathcal{B}_2$ on the order of a hundred pixels.

Given a triple of local features $\alpha_i, \alpha_j, \alpha_k$ at locations y_i, y_j, y_k on the reference grid, the steps of the search are the following

1. Precompute the locations of all local features in the image.

2. Assume N instances of local feature α_i in the image: $x_{i,1}, \ldots, x_{i,N}$.

3. For $n = 1, \ldots, N$, find all instances of α_j in $x_{i,n} + \mathcal{B}_1$; call these $x_{j,1}, \ldots, x_{j,M}$ (M may be 0).

   - For $m = 1, \ldots, M$, define $R_{x_{j,m} - x_{i,n}}$ to be the rotation determined by the vector $x_{j,m} - x_{i,n}$. For each instance of α_k at $x_k \in x_{i,n} + R_{x_{j,m} - x_{i,n}} \mathcal{B}_2$, determine the affine map T taking y_i, y_j, y_k into $x_{i,n}, x_{j,m}, x_k$.
   - Add (Tz_1, Tz_2, Tz_3) to the list of candidate bases.

An important constraint is that the size of the regions $\mathcal{B}_1, \mathcal{B}_2$ used in the image-wide search for the global arrangements be sufficiently large to guarantee that "coverage" at the reference pose extends to coverage in global coordinates (see Section 8.3 below). Specifically, we demand that if the registered ROI of a basis has at least three local features $\alpha_i, \alpha_j, \alpha_k$ somewhere in their distinguished neighborhoods in the reference grid, then this ROI will in fact be "hit" in the sense of finding an instance of the corresponding global arrangement in the original image coordinates. This is accomplished by choosing the size of the regions $\mathcal{B}_1, \mathcal{B}_2$ to be on the order of one hundred pixels. Specifically, in our applications it was sufficient to take $\mathcal{B}_1$ at most 11×11 (to accommodate the required range of scales) and $\mathcal{B}_2$ at most 7×7.

7 Final Classification

Final classification means assigning the label "object" or "background" to each candidate basis. This final disambiguation might be more computationally intensive than selection; this was our experience with detecting faces. One reason is that final classification generally requires both geometric and grey level *image normalization* whereas visual selection does not, at least not in our scheme. In our experiments, geometric normalization means registering the ROI around the basis to the reference grid and grayscale normalization means standardizing the registered intensity data. Similar techniques have been used elsewhere. After normalization, one typically computes a

fixed-length feature vector and classifies the candidates based on standard inductive methods (e.g., neural networks). The training set contains both "positive" examples from the object class and "negative" examples, which might be false positives from the selection stage. In our case we use regions-of-interest which are flagged by the triangle search in the types of generic images mentioned earlier.

We use classification trees for the final step. We recursively partition registered and standardized *edge* data. For each location in the reference grid we have four binary variables indicating the presence of one of the four edges in a 3×3 neighborhood of that point. When a candidate basis is detected, the associated affine transformation maps the locations of the edges in the ROI of the candidate basis into the reference grid, yielding a binary feature vector with one component for each of the 4 types of edges and each pixel in the reference grid. Several tens of trees are grown and aggregated as in (Amit & Geman 1997). The use of multiple trees together with photometrically invariant edge features provides a robust classifier.

Visual selection - the search for the global arrangements - together with final classification stage is therefore highly coarse-to-fine. One way to see this is that the organization of each step is tree-structured. For example, the edge fragments are defined as conjunctions of comparisons of intensity differences, organized as a vine; the search is terminated as soon as one comparison fails. Similarly, the point process determined by a local grouping is a thinning of the point process corresponding to the central edge; if the second edge is not found in the subregion determined by the central one (see Figure 2), the search is abandoned, and so forth. Finally, the global arrangements are strictly scarcer than the constituent local groupings and this search also has an underlying tree structure. This explains why the spatial distribution of processing illustrated in Figure 1 is so asymmetric. In contrast, if a neural network is trained to detect faces at a reference scale and then applied to every (or many) subregions of the image, the corresponding distribution would be more or less flat.

8 Background Densities and Parameter Selection

In this section we present some empirical results on the "statistics" of the local features defined above in generic images obtained from the web. These results guide the choice of parameters in order to obtain conditions **LIII**, **GI**, **GII**, which remain to be verified.

8.1 Density of Local Features

The "background density" of local features was estimated from 70 images randomly downloaded from the web. The local features were chosen by varying the number of edges N_{edges} (from $2-7$) and the size of the subregions N_{pixels} (from $7-40$) and using different shapes for the subregions. For each local feature we calculated the *density per pixel*, denoted λ_{local}, in each of the 70 image and computed the average, $\bar{\lambda}_{local}$, over images. We then regressed the *log* density on N_{edges} and N_{pixels}, obtaining

$$\log \bar{\lambda}_{local} = -5.96 - .64 N_{edges} + .15 N_{pixels} \tag{1}$$

with R^2 of 95%. It follows that, even at relatively close distances, the dependence among the individual edge fragments is sufficiently weak that if N_{pixels} is held fixed, the density itself scales like $(e^{-0.64})^{N_{edges}} \approx (0.5)^{N_{edges}}$. In particular, property **LIII** (low background density) is clearly satisfied in the ranges of parameters presented in Section 8.3 below.

Despite the high correlation, which is due to the averaging over images, there is substantial variation in the density from image to image. On the natural log-scale this variation is of order of ± 1. In Table 1 we display the mean and standard deviation of the log-density for $N_{pixels} = 10$ pixels for various values of N_{edges}. The value $N_{edges} = 0$ corresponds to the density of each of the four edges.

N_{edges}	0	1	2	3	4	5	6
mean	-3.8	-4.7	-5.2	-5.7	-6.3	-6.9	-7.5
std	.65	.83	.87	.97	.95	.93	.92

Table 1: Mean and standard deviation of local feature log-density over 70 random images for various values of N_{edges}, with $N_{pixels} = 10$

8.2 Density of Triangles

Consider again a triangle based on three local groupings $\alpha_0, \alpha_1, \alpha_2$. We used the 70 images to determine typical triangle densities in real images over a wide range of sizes for $\mathcal{B}_1, \mathcal{B}_2$ and offsets v_1, v_2 (triangle shapes). We searched for all instances of each triangle in each image. The density of the global arrangements can be predicted rather well from the density of the local features. If the three point processes defined by $\alpha_0, \alpha_1, \alpha_2$ were actually Poisson, each with the same density λ_{local}, and if these processes were mutually independent, then the density of the corresponding triangle would be

$$\lambda_{global} = \lambda_{local}^3 \cdot |\mathcal{B}_1| \cdot |\mathcal{B}_2|, \tag{2}$$

assuming we ignore small clustering effects. *In fact, the observed density of the triangles nearly obeys this equation.* In an additional test we replaced the exponent 3 in the expression for λ_{global} by a parameter η and estimated η by maximum likelihood based on the *counts* of the global arrangements. The maximum is very close to $\eta = 3$ with negligible variance.

Still, there are important exceptions to this seemingly straightforward Poisson analogy. For example, if α_0 and α_1 are both horizontal groupings of horizontal edges, and if v_1 respects this orientation, then *long range correlations* become significant and affect the estimates given above. Thus, knowing the local densities and given the near-Poisson nature of the corresponding point processes, one can obtain reasonable upper bounds on the densities of the global arrangements in generic scenes.

8.3 Choosing the Parameters

In order to estimate the likelihood of a missed detection, and thereby guide the choice of parameters, we need to estimate the probability that a registered object does not have any of the triangles (with the vertices in their distinguished neighborhoods). This is equivalent to having less than three of the local features at the specified locations. Recall that in training we kept only those local features which were over some threshold τ. *Assuming independence of these features on registered data*, and assuming the different fractions are approximately equal, we determine the false negative probability by a simple calculation using the binomial distribution. We can then choose N_{types}, the number of local features, in order to acquire the "coverage property" **GI** and maintain an acceptable level of error. We note that these estimates only require a small amount of training data since only the frequencies of local features are compiled and a degree of invariance is built in.

We calculated the frequencies of the special local features identified for faces in a training set of 300 faces as a function of N_{edges} and N_{pixels}. For these "common" local groupings, there is a strong linear relation with the number N_{edges} of edges and the size of the regions, N_{pixels}. The regression yielded $freq = .57 - .09N_{edges} + .03N_{pixels}$, with $R^2 = 93\%$. (Similar behavior is observed for randomly deformed latex symbols.) Choosing $N_{edges} = 3$ and $N_{pixels} = 10$ yields frequencies on the order of 50% which in turn leads to very low false negative rates with only order $N_{types} = 10$ local features; these are the values used in the experiments reported in the following section as well as in in (Amit et al. 1998). Clearly the local variability of the object class is crucial in determining these frequencies. However, it is not unrealistic to assume that, after factoring out linear variability, there are a good number of local groupings which appear in approximately 50% of the object images, near a fixed location of the reference grid.

With these choices for N_{edges}, N_{pixels} and N_{types}, the density λ_{local} of the local features is then order 10^{-3}. It follows from equation (2) that the density λ_{global} of the

global arrangements is order 10^{-5}. Since there are 120 model triangles, the density of detected global arrangements (and hence candidate bases) is order $120 \times 10^{-5} \sim 10^{-3}$, or approximately several tens per 100^2 pixels. Thus we see that, indeed, the conjunctions are very rare events in the background population, which is property **GII**.

In summary, it is possible to choose the parameters in order to achieve specific constraints on false alarms, missed detections and computation time. Of course there are the usual tradeoffs. For example, if N_{edges} and N_{pixels} are held fixed, then increasing N_{types} increases the number of false alarms but decreases the false negative rate, and similarly for N_{pixels}.

9 Experiments

The selection of candidate bases is determined by an image-wide search for the particular global arrangements which represent the object class, as discussed above.

In Figure 6 we show detection experiments including both visual selection and final classification, for the LaTeX symbols & and $\mathcal{Z}$ and for faces. The two symbol detectors are trained with 32 samples. The test images are 250x250 artificial scenes which contain 100 randomly chosen and randomly placed symbols in addition to the target one. The negative training examples were extracted from real scenes not the artificial scenes illustrated in Figure 6; consequently, the detection algorithm is independent of the particular statistics or other properties of these synthetic backgrounds. The lefthand panels of Figure 6 show all bases detected in the selection phase. Observe that a basis represents a precise hypothesis regarding the pose of the object. Processing time is approximately 20 seconds on a 166Mhz laptop pentium and 3 seconds on a Sparc 20.

For faces we trained on 300 pictures of 30 people (10 images per person) taken from the Olivetti database. The algorithm was tested on images from (Rowley et

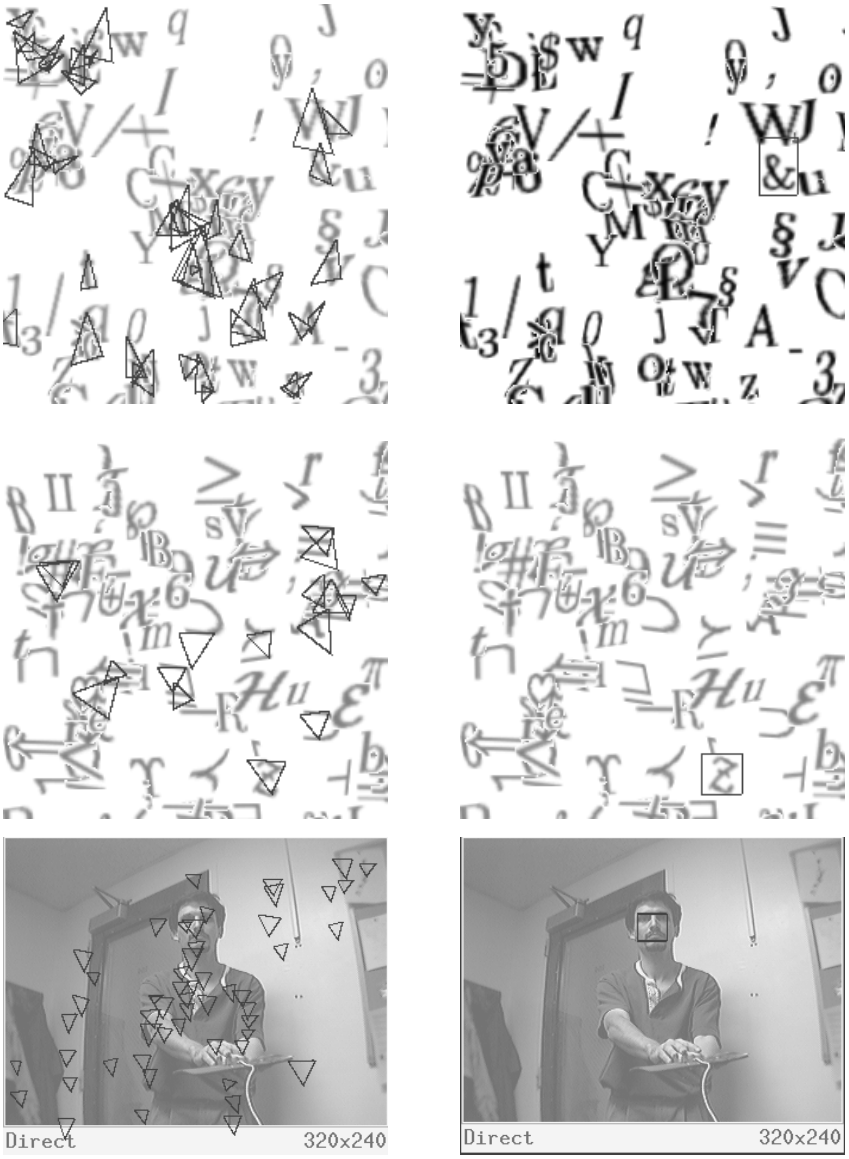

Figure 6: Top Left: All bases flagged by the &-detector. Top Right: Final decision. Middle - same thing for a $\mathcal{Z}$ detector. Bottom - same thing for the face detector.

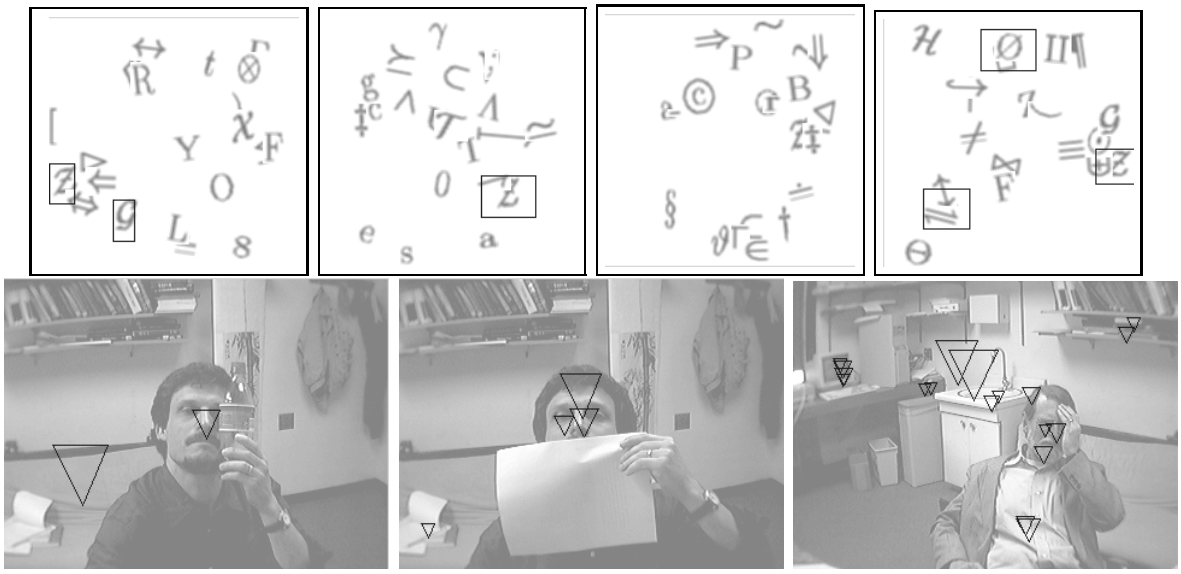

Figure 7: Top row: Experiments with occluded $\mathcal{Z}$'s. Bottom row: Experiments with occluded faces. The face is found during selection in all three images, but only retained during final classification in the lefthand one.

al. 1998) (for example Figure 1), and images captured on the Sun Videocam (for example Figure 6). Processing time on a Sparc 20 is approximately .5 seconds per 100×100 subimage. All computation times reported include six applications of the algorithm at different resolutions obtained by downsampling the original image by factors ranging from 1 (original resolution) to 1/4. About half of the processing time is spent in detecting the edges and the local groupings. Both operations are highly parallelizable.

In hundreds of experiments using pictures obtained from the videocam and Rowley's (Rowley et al. 1998) database the false negative rate of the visual selection stage is close to zero. Note that the visual selection part of the algorithm is inherently robust to partial occlusion. Since only three of the model features need to be found, the object is still detected if parts of it are degraded or occluded. It is hard to quantify these statements; however, in Figure 7 we show some results.

Some faces are lost during final classification. The main reason seems to be that

the final classifier is still trained using the 300 faces in the Olivetti training. This is a rather homogeneous data set in terms of lighting conditions, and other characteristics. One would need a larger number of examples of faces to improve the performance of this stage. Numerous results can be found at 'http://galton.uchicago.edu/∼amit/faces'.

10 Biological Vision

Our model was not conceived to explain how real brains function, although we have borrowed terms like "visual selection" and "foveation" from physiological and psychological studies in which these aspects of visual processing are well-established. In particular, there is evidence that object detection occurs in two phases - first searching for distinguished locations in a rather large field of view and then "focusing" the processing at these places. In this section we investigate some compelling links between our computational model and work on biological vision. We also consider an implementation using the architecture of artificial neural networks.

We have assumed that the only source of information for visual selection is grey level values from a single image; there is no color, motion or depth data. In other words, the procedure is entirely shape-based. It is obvious on empirical grounds that human beings analyze scenes without these additional cues. In addition, there are experiments in neuropsychology (e.g., (Bulthoff & Edelman 1992)) which indicate that 3D information is not crucial.

Our selection model has three clearly distinct levels of computation:

- Level I - edge fragments;
- Level II - local groupings of fragments;
- Level III - global arrangements of local groupings.

Level I roughly corresponds to the basic type of processing believed to be performed in certain layers of V1 (Hubel 1988). Level II involves more complex operations which

might relate to processing occurring in V2; and Level III could relate to functions of neurons in IT. These connections are elaborated in the next two subsections.

10.1 Flexible Groupings and Illusory Contours

How regular are the grey level patterns which activate cells in the brain? There is evidence of cells in various areas which respond to rather general stimuli. For example, in V1 there are responses to edge-like patterns which are orientation-dependent but contrast- independent (Schiller, Finlay & Volman 1976). And in (von der Heydt 1995) there is a review of the neurophysiological evidence for V2 cells responsive to "illusory" or "anomalous" contours; even in V1 according to (Grosof, Shapley & Hawken 1993). These cells respond equally well to an oriented line and to an occluded or interrupted line. They also respond to gradings which form the preferred orientation. Finally, cells in IT also respond to loose patterns and even to configurations which are difficult to name (Fujita, Tanaka, Ito & Cheng 1992). One interpretation of these experiments is that these cells respond to a *flexible* local configuration of edges constrained by loose geometrical relationships. Activation does not require a complete, continuous contour at a certain orientation; sufficient evidence for the presence of such a contour is enough.

This approach seems to be more robust and efficient than a finely-tuned search. Consider image contours arising from object boundaries and discontinuities in depth, lighting or shape. Such contours are often partially occluded or degraded by noise and therefore continuous contours may not be sufficiently stable for visual selection. Moreover, given that one observes a several nearby edge fragments of a certain orientation, it appears wasteful to attempt to "fill in" missing fragments and form a more complete entity. Since objects and "clutter" are locally indistinguishable, the additional information gain might be small compared, say, to inspecting another region. More specifically, detecting three approximately colinear horizontal edges in close proximity might be a rather unlikely event at a random image location, and

hence might sharply increase the likelihood of some non-accidental structure, such as an object of interest. However, conditioned on the presence of these three edge fragments, and on the presence of *either* an object *or* clutter, the remaining fragments needed to complete the contour might be very likely to be detected (or very unlikely due to occlusion) and hence of little use in discrimination. The fact that the visual system at the very low levels of LGN responds to contrast and not to homogeneous regions of lighting is another manifestation of the same phenomenon. Finally, the computation of these flexible groupings is local and it is not difficult to imagine a simple feed-forward architecture for detecting them from edge fragment data.

10.2 Global Arrangements and Invariance

There is clear evidence for translation and scale invariance within certain ranges in the responses of some neurons in IT, (Lueschow et al. 1994, Ito, Tamura, Fujita & Tanaka 1995). Most of these neurons do not select highly specific shapes. This is demonstrated in the experiments in (Kobatake & Tanaka 1994) and in (Ito et al. 1995) where successive simplifications of the selective stimuli, and various deformations or degradations, still evoke a strong response. Moreover the time between the local processing in V1 and the responses in IT, which involve integrating information in a rather large field of view and at a large range of scales, is a few tens of milliseconds.

Suppose a neuron in IT responds to stimuli similar to the types of global arrangements discussed here, and anywhere in the receptive field and over a range of scales. Then the speed of the calculation is at least partially explained by the simplicity of the structure it is detecting, which is not really an object but rather a more general structure, perhaps dedicated to many shapes simultaneously. However, conditioned on the presence of this structure, the likelihood of finding an object of interest in its immediate vicinity is considerably higher than at a random location.

Put another way, the neurons in IT seem to have already overcome the problem of "moding out" scale, translation and other types of deformations and degradations.

This would appear to be very difficult based on complex object representations. It is more efficient to use sparse representations for which it is easy to define those *disjunctions* needed for invariance. Scale and deformation invariance are achieved by taking disjunctions over the angles and distances between the local features; occlusion and degradation invariance are achieved by taking a disjunction over several spatial arrangements (the different triangles).

10.3 Segmentation

There is no segmentation in the sense of a processing stage which precedes recognition and extracts a rough approximation of the bounding contours of the object. The classical "bottom-up" model of visual processing assumes that edge information leads to the segmentation of objects. This is partly motivated by the widespread assumption that local processing carried out in V1 involves the detection, and possibly organization, of oriented edge segments (Hubel & Wiesel 1977, Hubel 1988). However, edge detectors do not directly determine smooth, connected curves which delineate well-defined regions and it is now clear to many researchers in both computer and biological vision that purely edge-based segmentation is not feasible in most real scenes (von der Heydt 1995, Ullman 1996), at least not without a tentative interpretation of the visual input.

10.4 Architecture

Our actual implementation of the visual selection algorithm is of course entirely serial. However, suppose we consider the type of multi-layer arrays of processors which are common in neural models and suppose a large degree of connectivity. Then what sort of architecture might be efficient for the detection of the types of global arrangements we have described? In particular, how would one achieve invariance to scale, translation and other transformations with a reasonable number of units and connections?

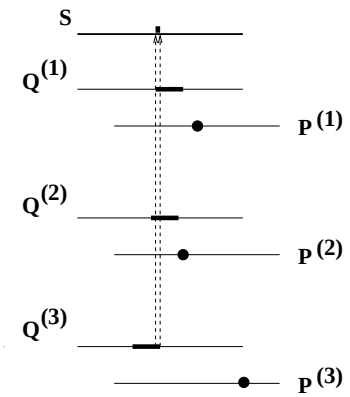

Figure 8: The $P^{(i)}$ arrays detect the local features. The dots in the P arrays are points where the corresponding local feature was found. The thick lines in the Q arrays are the locations activated due to activity in the associated P arrays. The widths of these lines correspond to the sizes of the boxes $\mathcal{B}_i$. Finally the thick dot in the S array shows where activation occurs due to the presence of a sufficient number (3) of active Q arrays.

First it is clear that the edges and local features are easily detected in a parallel architecture with local processing. "Virtual centers" of global arrangements can also be detected using a parallel architecture. The price is loss of some pose information. In other words the object is detected over the range of poses, but the detection is represented only through the center and hence information on scale and rotation is lost. The idea is the following. For each local features $\alpha_i, i = 1 \ldots, N_{types}$, at location y_i, we determine a region of variation B_i relative to the center of the reference grid, which accommodates the expected variations in scale, rotation etc. The constraints on each of the points relative to the center are now *decoupled*. Each local feature α_i that is found in a detection array $P^{(i)}$ say at x, activates all the locations in the region $x - \mathcal{B}_i$ in an auxiliary array $Q^{(i)}$. These are all the locations of an object center which could produce a feature α_i at x, if an object was present there within the allowed range of poses.

The activities in the auxiliary arrays are summed into an array S and those locations which exceed some threshold are taken as candidate object centers. This is

precisely a parallel implementation of the generalized Hough transform. The detected locations are represented through activities in the retinotopic layer S. A diagram illustrating this architecture is presented in Figure 8. Note that this network is dedicated to a specific object representation, i.e. a specific list of local features and locations. In (Amit 1998) we show how a *fixed* architecture with a moderate number of arrays can accommodate *any* detection task with a central memory module storing the representations of the various objects.

10.5 Multiple Object Classes

Remarkably, real brains manage to parse full scenes and perform rapid visual selection when *no specific detection task is specified*, i.e., no prior information is provided about objects of interest. Clearly at least thousands of possible object classes are then simultaneously considered. Perhaps context plays a significant role; see (Biederman 1981) and (Palmer 1975).

More modestly, how might a computer algorithm be designed to conduct an efficient search for say tens or hundreds of object classes? Ideally, this would be done in some coarse-to-fine manner, in which many object classes are *simultaneously* investigated, leading eventually to finely-tuned distinctions. Clearly, efficient indexing is crucial (Lowe 1985).

Although we have concentrated here on a single object class, it is evident that the representations obtained during training could be informative about many objects. Some evidence for this was discussed in (Amit & Geman 1997) in the context of "shape quantization"; decision trees induced from training data about one object class were found to be useful for classifying shapes never seen during training.

We are currently trying to represent multiple object classes by arrangements of local groupings in much the same manner as discussed in this paper for a single object class. The world of spatial relationships is exceptionally rich and our previous experience with symbol detection is promising. We expect the number of arrangements

needed to identify multiple classes, or separate them from each other, will grow logarithmically with the number of classes. The natural progression is to first separate all objects of interest from "background" and then begin to separate object classes from one another, eventually arriving at very precise hypotheses. The organization of the computation is motivated by the "twenty questions paradigm"; the processing is tree-structured and computational efficiency is measured by mean path length.

11 Conclusion

The main strengths of the proposed model are stability, computational efficiency, and the relatively small amount of training data. For example, in regard to face detection, we have tested the algorithm under many imaging conditions, including "on-line" experiments involving a digital camera in which viewing angles and illumination vary considerably and objects can be partially occluded. It is likely that the algorithm could be accelerated to nearly real-time. One source of these properties is the use of crude, image-based features rather than refined, model- based features; any "sub-classification" problems are eliminated. Another source is the explicit treatment of photometric and geometric invariance. And finally there is the surprising uniformity of the "statistics" of these features in both object and background populations, which can be learned from a modest number of examples, and which determine error rates and total computation.

The main limitations involve accuracy and generality. First, there is a non-negligible false negative rate (e.g., five percent for faces) if the number of regions selected for final classification is of order 10-100. This is clearly well below human performance, although comparable to other detection algorithms. Second, we have not dealt with general poses or 3D aspects; whereas scale and location are arbitrary, we have by no means considered all possible viewing angles. Finally, our model is dedicated to a specific object class and does not account for general *scene parsing*.

How is visual selection guided when no specific detection task is required and a great many objects of interest, perhaps thousands, are simultaneously spotted?

Acknowledgement. The authors would like to thank Daniel Amit and the referees for many helpful comments.

References

Amit, Y. (1998), A neural network architecture for visual selection, Technical Report 474, University of Chicago.

Amit, Y. & Geman, D. (1997), 'Shape quantization and recognition with randomized trees', *Neural Computation* **9**, 1545–1588.

Amit, Y., Geman, D. & Jedynak, B. (1998), Efficient focusing and face detection, *in* H. Wechsler & J. Phillips, eds, 'Face Recognition: From Theory to Applications, NATO ASI Series F', Springer-Verlag, Berlin.

Biederman, I. (1981), On the semantic of a glance at a scene, *in* M. Kubovy & J. R. Pomerantz, eds, 'Perceptual Organization', Lawrence Erlbaum Assoc., Hillsdale, N.J.

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

Bulthoff, H. H. & Edelman, S. (1992), 'Psychophysical support for a two-dimensional view interpolation theory of object recognition', *Proc. Natl. Acad. Sci.* **89**, 60–64.

Desimone, R., Miller, E. K., Chelazzi, L. & Lueschow, A. (1995), Multiple memory systems in visual cortex, *in* M. S. Gazzaniga, ed., 'The Cognitive Neurosciences', MIT Press, Cambridge, Massachusetts, pp. 475–486.

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

Grosof, D. H., Shapley, R. M. & Hawken, M. J. (1993), 'Macaque v1 neurons can signal "illusory" contours', *Nature* **365**, 550–552.

Hubel, H. D. (1988), *Eye, Brain, and Vision*, Scientific American Library, New York.

Hubel, H. D. & Wiesel, T. N. (1977), 'Ferrier lecture: Functional architecture of macaque monkey visual cortex', *Proc. Roy. Soc. Lond. [Biol.]* **98**, 1–59.

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.

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.

Lamdan, Y., Schwartz, J. T. & Wolfson, H. J. (1988), Object recognition by affine invariant matching, *in* 'Proc. IEEE Conf. on Computer Vision and Pattern Recognition', pp. 335–344.

Lowe, D. G. (1985), *Perceptual Organization and Visual Recognition*, Kluwer Academic Press, Boston.

Lueschow, A., Miller, E. K. & Desimone, R. (1994), 'Inferior temporal mechanisms for invariant object recognition', *Cerebral Cortex* **5**, 523–531.

Maurer, T. & von der Malsburg, C. (1996), Tracking and learning graphs and pose on image sequences of faces, *in* 'Proceedings, Second International Conference on Automatic Face and Gesture Recognition', IEEE Computer Society Press, pp. 176–181.

Palmer, S. E. (1975), 'The effects of contextual scenes on the identification of objects', *Memory and Cognition* **3**, 519–526.

Rojer, A. S. & Schwartz, E. L. (1992), A quotient space hough transform for scpae-variant visual attention, *in* G. A. Carpenter & S. Grossberg, eds, 'Neural Networks for Vision and Image Processing', MIT Press.

Rowley, H. A., Baluja, S. & Takeo, K. (1998), 'Neural network-based face detection', *IEEE Trans. PAMI* **20**, 23–38.

Schiller, P., Finlay, B. L. & Volman, S. F. (1976), 'Quantitative studies of single-cell in monkey striate cortex. i. spatiotemporal organization of receptive fields', *Journal of Neurophysiology* **39**, 1288–1319.

Sung, K. K. & Poggio, T. (1998), 'Example-based learning for view-based face detection', *IEEE Trans. PAMI* **20**, 39–51.

Thorpe, S., Fize, D. & Marlot, C. (1996), 'Speed of processing in the human visual system', *Nature* **381**, 520–522.

Ullman, S. (1996), *High-Level Vision*, M.I.T. Press, Cambridge, MA.

Van Essen, D. C. & Deyoe, E. A. (1995), Concurrent processing in the primate visual cortex, *in* M. S. Gazzaniga, ed., 'The Cognitive Neurosciences', MIT Press, Cambridge, Massachusetts, pp. 475–486.

von der Heydt, R. (1995), Form analysis in visual cortex, *in* M. S. Gazzaniga, ed., 'The Cognitive Neurosciences', MIT Press, Cambridge, Massachusetts, pp. 365–382.

Winston, P. H. (1970), *Learning Structural Descriptions from Examples*, Doctoral Dissertation, M.I.T.