Communicated by Dana Ballard

Part Segmentation for Object Recognition

Alex Pentland
Vision Sciences Group, The Media Lab, Massachusetts Institute of Technology, Room E15-410, 20 Ames Street, Cambridge, MA 02139, USA

Visual object recognition is a difficult problem that has been solved by biological visual systems. An approach to object recognition is described in which the image is segmented into parts using two simple, biologically-plausible mechanisms: a filtering operation to produce a large set of potential object "parts," followed by a new type of network that searches among these part hypotheses to produce the simplest, most likely description of the image's part structure.

1 Introduction

In order to recognize objects one must be able to compute a stable, canonical representation that can be used to index into memory (Binford 1971; Marr and Nishihara 1978; Hoffman and Richards 1985). The most widely accepted theory on how people recognize objects seems to be that they first segment the object into its component parts and then recognition occurs by using this part description to classify the object, perhaps by use of an associative network.

Despite the importance of object recognition, most vision research — and especially neural network research — has been aimed at understanding early visual processing. In part this focus on early vision is because the uniform, parallel operations typical of early vision are easily mapped onto neural networks, and are more easily understood than the nonhomogeneous, nonlinear processing required to segment an object into parts and then recognize it. As a consequence, the process of object recognition is little understood.

The goal of this research is to automatically recover accurate part descriptions for object recognition. I have approached this objective by developing a system that segments an imaged object into convex parts using a neural network that is similar to that described by Hopfield and Tank (Hopfield and Tank 1985), but which uses a temporally-decaying feedback loop to achieve considerably better performance. For the sake of efficiency and simplicity I have used silhouettes, obtained from grey-scale images by intensity, motion, and texture thresholding, rather than operating on the grey-scale images directly.

2 A Computational Theory of Segmentation

Many machine vision systems employ matched filters to find particular 2-D shapes in an image, typically using a multiresolution approach that allows efficient search over a wide range of scales. Thus, in machine vision, a natural way to locate the parts of a silhouetted object is to make filter patterns that cover the spectrum of possible 2-D part-shapes (as is shown in figure 1(a)), match these 2-D patterns against the silhouette, and then pick the best matching filter. If the match is sufficiently good, then we register the detection of a part whose shape is roughly that of the best-matching filter.

A biological version of this approach might use many hypercolumns each containing receptive fields with excitatory regions shaped as in figure 1. The cell with the best-matching excitatory field would be selected by introducing strong lateral inhibition within the hypercolumn in order to suppress all but the best-responding cells. This arrangement of receptive fields and within-hypercolumn inhibition produces receptive fields with oriented, center-surround spatial structure, such as is shown in figure 1(b).

The major problem with such a filtering/receptive field approach is that all such techniques incorporate a noise threshold that balances the number of false detections against the number of missed targets. Thus we will either miss many of the object's parts because they don't quite fit any of our 2-D patterns, or we will have a large number of false detections.

This false-alarm versus miss problem occurs in almost every image processing domain, and there are only two general approaches to overcoming the problem. The first is to improve the discriminating power of the filter so as to improve the false-alarm/miss tradeoff. The success of this approach depends upon precise characterization of the target and so is not applicable to this problem.

In the second approach, each non-zero response of a filter/receptive field is considered as an *hypothesis* about the object's part structure rather than being considered as a detection. One therefore uses a very low threshold to obtain a large number of hypotheses, and then searches through them to find the "real" detections. This approach depends upon having some method of measuring the likelihood of a *set* of hypotheses, i.e., of measuring how good a particular segmentation into parts is as an explanation of the image data. It is this second, "best explanation" approach that I have adopted in this paper.

2.1 Global Optimization: The Likelihood Principle and Occam's Razor. The notion that vision problems can be solved by optimizing some "goodness of fit" measure is perhaps the most powerful paradigm found in current computational research (Hopfield and Tank 1985; Ballard

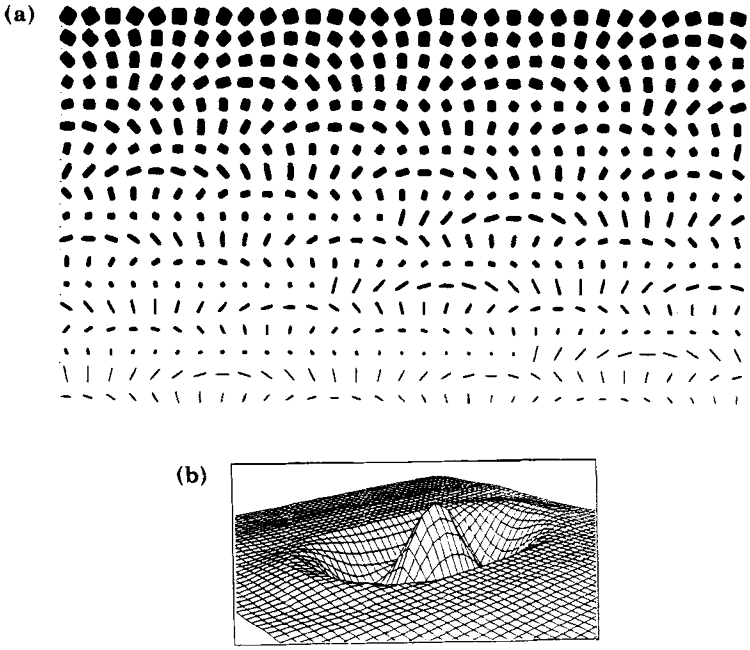

Figure 1: (a) Two-dimensional binary patterns used to segment silhouettes into parts. (b) Spatial structure of a receptive field corresponding to one of these binary patterns.

et al. 1983; Hummel and Zucker 1983; Poggio et al. 1985). Although heuristic measures are sometimes employed, the most attractive schemes have been based on the likelihood principle (the scientific principle that the most likely hypothesis is the best one), i.e., they have posed the problem in terms of an *a priori* model with unknown parameter values, and then searched for the parameter settings that maximize the likelihood of the model given the image data.

Recently it has been proven (Rissanen 1983) that one method of finding this maximum likelihood estimate is by use of the formal, information-theoretic version of Occam's Razor: the scientific principle that the *simplest* hypothesis is the best one. In information theory the simplicity or complexity of a description is measured by the number of bits (binary digits) needed to encode both the description and remaining residual noise. This new result tells us that both the likelihood principle and Occam's Razor agree that the *best* description of image data is the one that provides the bitwise shortest encoding.

This method of finding the maximum likelihood estimate is particularly useful in vision problems because it gives us a simple way to produce maximum likelihood estimates using image models that are too complex for direct optimization (Leclerc 1988). In particular, to find the maximum likelihood estimate of an object's part structure one needs only to find the shortest description of the image data in terms of parts.

2.2 A Computational Procedure. How can the shortest/most likely image description be computed? Let $\{H\}$ be a set of n part hypotheses h_i produced by our filters/receptive fields, and let $\{H^*\}$ be a subset of $\{H\}$ containing m hypotheses. The particular elements which comprise $\{H^*\}$ can be indicated by a vector $\vec{x}$ consisting of $n - m$ zeros and m ones, with a one in slot i indicating that hypothesis h_i is an element of $\{H^*\}$.

The presence of part hypothesis h_i in the set $\{H\}$ indicates that a particular pattern from among those illustrated in figure 1(a) has at least a minimal correspondence to the image data at some particular image location. Let us designate the number of image pixels at which h_i and the image agree (have the same value) by a_{ii}, and the number of image pixels at which h_i and the image disagree (have different values) by e_{ii}. Then h_i provides an *encoding* of the image which saves $S(h_i)$ bits as compared to simple pixel-by-pixel description of the image pixel values. The amount of this savings, in bits, is:

$$S(h_i) = k_1 a_{ii} - k_2 e_{ii} - k_3. \tag{2.1}$$

where k_1 is the average number of bits needed to specify a single image pixel value, k_2 is the average number of bits needed to specify that a particular pixel is erroneously encoded by h_i, and k_3 is the cost of specifying h_i itself. The ratio between k_1 and k_2 is our *a priori* estimate of the signal to noise ratio, including both image noise and noise from quantization of the set of 2-D shape patterns. The parameter k_3 is equal to the minus log of the probability of a particular part hypothesis. By default we make k_3 equal for all h_i; however, we can easily incorporate *a priori* knowledge about the likelihood of each h_i by setting k_3 to the minus log probability associated with each h_i.

Equation 2.1 allows us to find the *single* hypothesis which provides the best image description by simply maximizing $S(h_i)$ over all the hypotheses h_i. To find the overall maximum-likelihood/simplest description, however, we must search from among the power set of $\{H\}$ to find that subset $\{H^*\}$ which maximizes $S(\vec{x})$. Thus we must be able to account for interactions between the various h_i in $\{H^*\}$.

Let a_{ij} be the number of image pixels at which h_i, h_j, and the image all agree, and e_{ij} be number of image pixels at which both h_i and h_j disagree with the image. We then define a matrix **A** with values a_{ii} on the diagonal, and values $-1/2a_{ij}$ for $i \neq j$, and similarly a matrix **E** with values e_{ii} on the diagonal, and values $-1/2e_{ij}$ for $i \neq j$. Ignoring points where three or more h_i overlap, the savings generated by encoding the image data using $\{H^*\}$ (as specified by the vector $\vec{x}$) is simply

$$S(\vec{x}) = k_1\vec{x}\mathbf{A}\vec{x}^T - k_2\vec{x}\mathbf{E}\vec{x}^T - k_3\vec{x}\vec{x}^T. \tag{2.2}$$

Equation 2.2 can easily be extended to include overlaps between three or more parts by adding in additional terms that express these higher-order overlaps. However, these higher-order overlaps are expensive to calculate. Moreover, such high-order overlaps seem to be infrequent in real imagery. I have chosen, therefore, to assume that in the *final* solution that there are a negligible number of image points covered by three or more h_i. Note that we are *not* assuming that this is true of the entire set $\{H\}$, where such high-order overlaps will be common. The important consequence of this assumption is that the maximum of the savings function $S(\vec{x})$ over all $\vec{x}$ is also the maximum of equation 2.2.

The solution to equation 2.2 is straightforward when the matrix **Q**

$$\mathbf{Q} = k_1\mathbf{A} - k_2\mathbf{E} - k_3\mathbf{I} \tag{2.3}$$

is positive (or negative) definite. Unfortunately, this is not the case in this problem. As a consequence, relaxation techniques (Hummel and Zucker 1983) such as the Hopfield-Tank network (Hopfield and Tank 1985) give a very poor solution.

I have therefore devised a new method of solution (and corresponding network) which *can* provide a good solution to equation 2.2. This new technique is a type of *continuation method*: one first picks a problem related to the original problem that *can* be solved, and then iteratively solves a series of problems that are progressively closer to the original problem, each time using the last solution as the starting point for the next iteration.

In the problem at hand, **Q** is easily solved when k_3 is large enough, as then **Q** is diagonally dominant and thus negative definite. Therefore, I can obtain a *globally* good solution by first solving using a large k_3, and then — using that answer as starting point — progressively resolve using smaller and smaller values of k_3 until the desired solution is obtained. Because k_3 is the cost of adding a model to our description, the effect of this continuation technique is to solve for the largest, most prominent parts first, and then to progressively add in smaller and smaller parts until the entire figure is accounted for.

The neural network interpretation of this solution method is a Hopfield-Tank network placed in a feedback loop where the diagonal weights are initially quite large and decay over time until they finally reach the desired values. In each "time step" the Hopfield-Tank network stabilizes, the diagonal weights are reduced, and the network outputs are fed back into the inputs. When the diagonal weights reach their final values, the desired outputs are obtained.

It can be shown that for many well-behaved problems (for example, when the largest eigenvalues are all of one sign, with opposite-signed eigenvalues of much smaller magnitude) this feedback technique will produce an answer that is on average substantially better than that obtained by Hopfield-Tank or relaxation methods. As with relaxation techniques (Hummel and Zucker 1983), this feedback method can be applied to problems with asymmetric weights.

A biological equivalent of our solution method is to use a set of hypercolumns (each containing cells with the excitatory subfields illustrated in figure 1) that are tied together by a Hopfield-Tank network augmented by a time-decaying feedback loop. The action of this network is to suppress activity in all but a small subset of the hypercolumns. After this network has stabilized, each of the remaining active cells correspond exactly to one part of the imaged object. The characteristics of that cell's excitatory subfield correspond to the shape of the imaged part.

3 Segmentation Examples

This technique has been tested on over two hundred synthetic images, with widely varying noise levels (Pentland 1988). In these tests the number of visible parts was correctly determined 85–95% of the time (depending on noise level), with largely obscured or very small parts accounting for almost all of the errors. Estimates of part shape were similarly accurate. The following three examples illustrate this segmentation performance.

The first example uses synthetic range data with a dynamic range of 4 bits. In this example, only 72 2-D shape patterns were employed in order to illustrate the effects of coarse quantization in both orientation and size. The intent of this example is to demonstrate that a high-quality segmentation into parts can be achieved despite coarse quantization in both orientation, size, and range values, and despite wide variation in the weights. In the remaining examples, the 2-D shape patterns shown in figure 1(a) were employed.

Figure 2(a) shows an intensity image of a CAD model; synthetic range data from this model is shown in figure 2(b). These range data were histogrammed and automatically thresholded, producing the silhouette shown in figure 2(c).

Figure 2(d) shows the operation of our new solution method. The parameter k_3 is initially set to a large value, thus making equation 2.2 diagonally dominant. In this first step only the very largest parts are recovered, as is shown in the first frame of figure 2(d). The parameter k_3 is then progressively reduced and the equation resolved, allowing smaller and smaller parts to be recovered. This is shown in the remaining frames of figure 2(d). This solution method therefore constructs a scale hierarchy of object parts, with the largest and most visible at the top of the hierarchy and the smallest parts on the bottom. This scale hierarchy can be useful in matching and recognition processes.

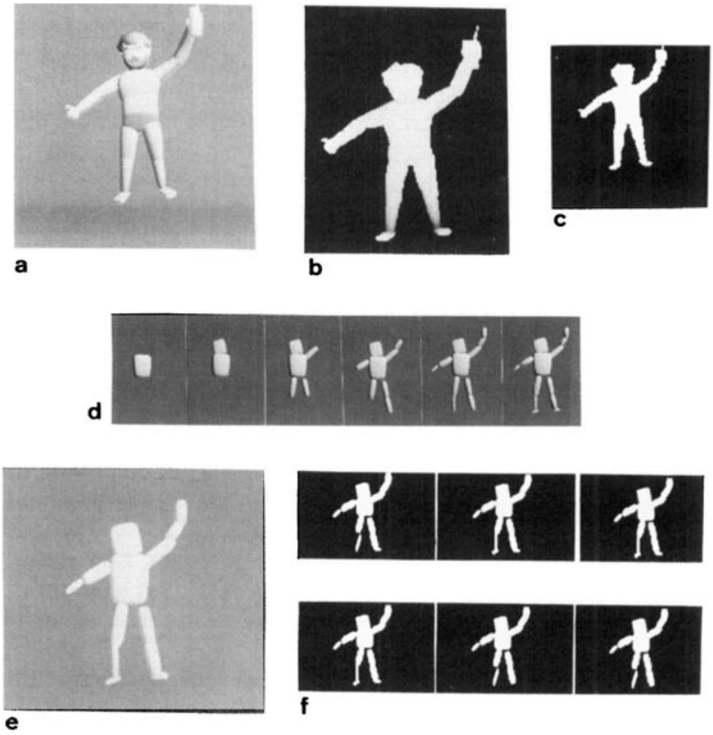

Figure 2: (a) Intensity image of a CAD model. (b) Range image of this model. (c) Silhouette of the range data. (d) This sequence of images illustrates how our continuation method constructs a scale-space description of part structure, first recovering only large, important parts and then recovering progressively smaller part structure. (e) Final segmentation into parts obtained using only very coarsely quantized 2-D patterns; 3-D models corresponding to recovered parts are used to illustrate the recovered structure. (f) Segmentations for a 5 : 1 ratio of the parameters k_i, showing that the segmentation is stable.

The final segmentation for this figure is shown in figure 2(e); here 3-D volumetric models have been substituted for their corresponding[1] 2-D shapes in order to better illustrate how the silhouette was segmented into parts. The z dimension of these 3-D models is arbitrarily set equal to the smaller of the x and y dimensions. It can be seen that, apart from coarse quantization in orientation and size, the part segmentation is a good one.

[1] That is, for each 2-D pattern we substituted a 3-D CAD model whose outline corresponds exactly to the 2-D shape pattern.

One important question is the stability of segmentation with respect to the parameters k_i. Figure 2(f) shows the results of varying the ratio of parameters k_1, k_2, and k_3 over a range of $5:1$. It can be seen that the part segmentation is stable, although as the relative cost of each model increases (the final value of k_3 becomes large) small details (such as the feet) disappear.

The second example of segmenting a silhouette into parts uses a real image of a person, shown in figure 3(a). A silhouette was produced by automatic thresholding of a fractal measure of texture smoothness; this silhouette is shown in figure 3(b). The resulting segmentation into parts is shown in figure 3(c).

An example of segmenting a more complex silhouette into parts uses the Rites of Spring, a drawing by Picasso, shown in figure 3(c). The area within the box was digitized and the intensity thresholded to produce a

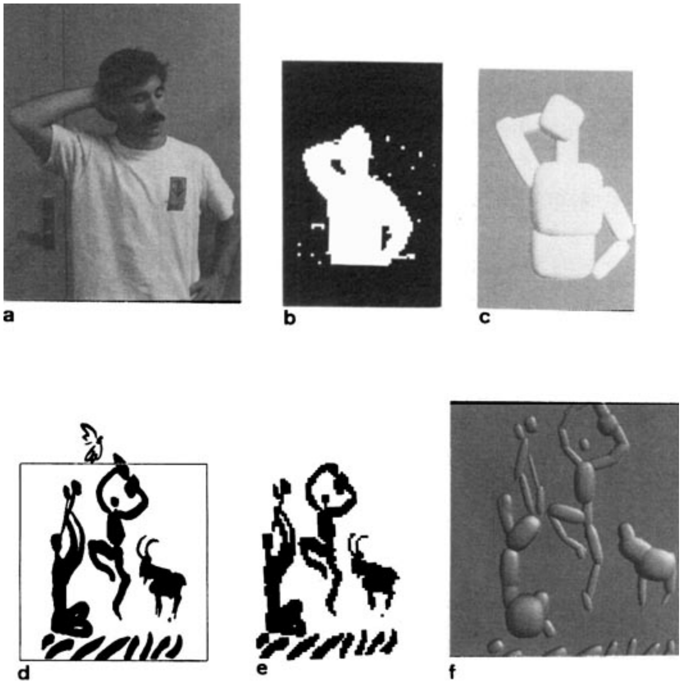

Figure 3: (a) Image of a person. (b) Silhouette produced by thresholding a fractal texture measure. (c) Automatic segmentation into parts. (d) The Rites of Spring, by Picasso. (e) Digitized version. (f) The automatic segmentation into parts.

coarse silhouette, as shown in figure 3(d). The automatic segmentation is shown in figure 3(e). It is surprising that such a good segmentation can be produced from this hand-drawn, coarsely digitized image (note that very small details, e.g., the goat's horns, were missed because they were smaller than any of the 2-D patterns).

4 Summary

I have described a method for segmenting 2-D images into their component parts, a critical stage of processing in many theories of object recognition. This method uses two stages: a detection stage which uses matched filters to extract hypotheses about part structure, and an optimization stage, where all hypotheses about the object's part structure are combined into a globally optimum (i.e., simplest, most likely) explanation of the image data. The first stage is implemented by local competition among the filters illustrated in figure 1(a), and the second stage is implemented by a new type of neural network that gives substantially better answers than previously suggested optimization networks. This new network may be described as a relaxation or Hopfield-Tank network augmented by time-decaying feedback. For additional details the reader is referred to reference (Pentland 1988).

Acknowledgments

This research was made possible in part by National Science Foundation, Grant No. IRI-8719920. I wish to especially thank Yvan Leclerc for his comments, and for reviving my interest in minimal length encoding.

References

Ballard, D.H., G.E. Hinton, and T.J. Sejnowski. 1983. Parallel Visual Computation. *Nature* **306**, 21–26.

Binford, T.O. 1971. Visual Perception by Computer. *Proceeding of the IEEE Conference on Systems and Control*, Miami.

Hoffman, D. and W. Richards. 1985. Parts of Recognition. *In:* From Pixels to Predicates, ed. A. Pentland. New Jersey: Ablex Publishing Co.

Hopfield, J.J. and D.W. Tank. 1985. Neural Computation of Decisions in Optimization Problems. *Biological Cybernetics* **52**, 141–152.

Hummel, R.A. and S.W. Zucker. 1983. On the Foundations of Relaxation Labeling Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **5:3**, 267–287.

Leclerc, Y. 1988. Construction Simple Stable Descriptions for Image Partitioning. *Proc. DARPA Image Understanding Workshop*, April 6–8, Boston, MA, 365–382.

Marr, D. and K. Nishihara. 1978. Representation and Recognition of the Spatial Organization of Three-dimensional Shapes. *Proceedings of the Royal Society–London B* **200**, 269–94

Pentland, A. 1988. *Automatic Recovery of Deformable Part Models.* Massachusetts Institute of Technology Media Lab Vision Sciences Technical Report 104.

Poggio, T., V. Torre, and C. Koch. 1985. Computational Vision and Regularization Theory. *Nature* **317**, 314–319.

Rissanen, J. 1983. Minimum-length Description Principle. *Encyclopedia of Statistical Sciences* **5**, 523–527. New York: Wiley.

Received 23 September; accepted 8 November 1988.