9

INTEGRATED DESCRIPTIONS FOR VISION

by

Trevor Darrell

B.S.E, Computer Science and Enginering
University of Pennsylvania, 1988

Submitted to the Media Arts and Sciences
Section, School of Architecture and Planning in partial
fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1990

Author ....
August 10, 1990

Certified by ....
Alex P. Pentland
Associate Professor
Thesis Supervisor

Accepted by ....
Stephen Benton
Chairman
Departmental Committee
on Graduate Studies

INTEGRATED DESCRIPTIONS FOR VISION

by

Trevor Darrell

Submitted to the Media Arts and Sciences
Section, School of Architecture and Planning
on August 10, 1990, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

We formulate the segmentation task as a search for a set of descriptions which minimally encodes a scene. Descriptions can represent many image formation processes, including shape functions, texture generators, and illumination models. We develop a *cooperative robust estimation* framework that can estimate the parameters and segmentation of a set of descriptions, together with a modified Hopfield-Tank network that finds the subset with the shortest encoding length.

Part-based image representations are a natural way to describe a wide range of natural and man-made objects, and are useful in describing many portions of common scenes. Our initial application of this segmentation paradigm has been to find the minimal part-based description of an image; we show the results of this decomposition on a variety of synthetic and real-world range images, as well as a few intensity images.

Thesis Supervisor: Alex P. Pentland
Title: Associate Professor

Contents

List of Figures

With very special thanks
for continued love and support,
this thesis is dedicated to
Mom and Dad!

Chapter 1

Introduction

Visual perception is the recovery of the simplest representation of an image which captures all the "essential" information in the scene. What constitutes "essential" for a given scene is, of course, dependent both on the intended task and on prior experience. This is evident in everyday experience: when I look on my desk, I first notice a cup-like shape filled with liquid, not a complicated assortment of intensities, colors and reflections. Somehow my visual system recovers the core information about shape, surface properties, and illumination from the raw visual image, which I can then recognize as my coffee mug.

We are interested in the mechanisms visual systems use to extract this core information. Since they must be defined in the context of prior experience, we use descriptive models that capture basic, oft-repeated patterns that exist in the world. If we can parse images into their constituent descriptions, we can generate compact representations of even complicated images. Most images are produced by a collection of processes in the scene which are individually straightforward to describe, so a small number of models combined in simple ways often suffices to represent an image.

Figure 1-1 shows an integrated description of a range image, using a "blob"-ish description model. Descriptions need not be an exact reconstruction of the original data; sometimes the salient information is captured in a caricature or smoothed version of the image. How to decide what the simplest integrated description is depends on what descriptions are available to be used; the best description is the shortest one possible in our descriptive language. We want the set of descriptions that constitutes the minimal encoding of the image, including both the cost

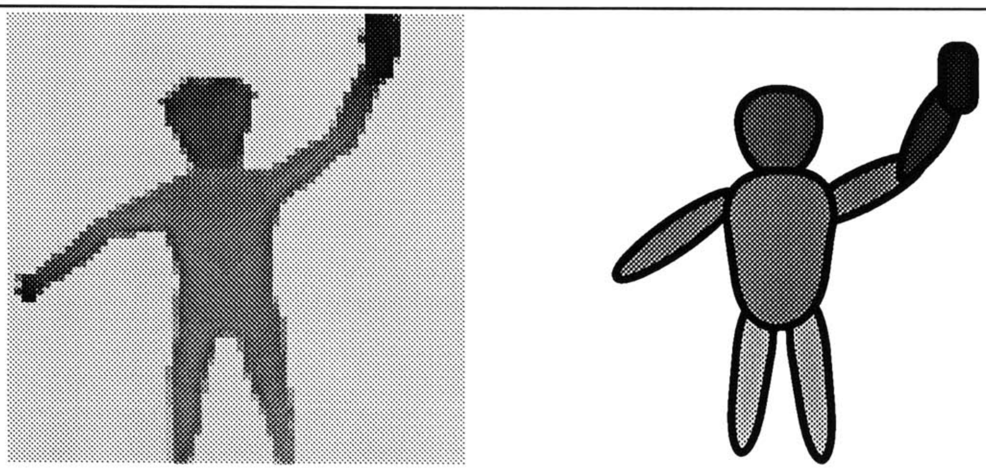

Figure 1-1: Synthetic range image and sample "blob"-based description

of encoding the model as well as the residual error.

Even if we know how to estimate and generate the individual description models, the task of finding an integrated set of descriptions is hardly straightforward; there is a critical dilemma between estimation and segmentation to be overcome:

- Estimating the parameters for a description depends directly on its segmentation, e.g. what portion of the image the description covers.
- Determining the correct segmentation of a description depends on its parameters and estimated data, but also on the other descriptions in the image.

This thesis presents a framework for the recovery of integrated image descriptions, using a architecture that mirrors this estimation/segmentation duality in an attempt to overcome it.

At first, we have no knowledge about actual scene structure, and can only begin with a collection of guesses of possible descriptions based on previous experience. The core of our recovery framework is an iteratiion, first estimating description parameters given a segmentation, and then eliminating unproductive descriptions in order to refine the segmentation estimates. Both the estimation of individual descriptions and the determination of which descriptions are "useful," are based on a minimization of encoding length.

The remainder of this thesis begins with a chapter discussing image descriptions and the use of minimal length encoding techniques. A parallel extension of robust estimation has been developed to estimate integrated descriptions, and is described in Chapter 2. Chapter 3 motivates the selection of a "good" subset of a description set using a network-based optimization. Finally, we consider the application of this framework to the task of "part-based" segmentation, outline a prototype implementation, and present some empirical results.

Chapter 2

Descriptions and Minimal Encoding

The question of how to evaluate a representation of an image is a fundamental concern of many fields. Several general criteria have been proposed:

- find the information in the signal (Shannon)
- find the most likely model of the data (Bayes)
- find the simplest representation (Gestalt)

Much of the work in vision research has been based on one of these principles. Despite their different origins, intuitively they seem to have converged on a single notion of what constitutes a good representation. Indeed, all of these approaches can be unified under a single framework, that of finding the representation that minimally encodes a image.

The first of these, the theory of information laid out by Claude Shannon, has motivated much of the contemporary work in image coding [7]. He defined "entropy" to be to the lack of predictability between elements in a representation; if there is some predictability from one element to another, then entropy is not at its maximum, and a shorter encoding can be constructed. When the encoding cannot be compressed further, the resulting signal consists of "pure information." Thus if we find the representation with the minimal encoding, in some sense we have found the information in the image.

Probabilistic inference has a long tradition in statistics as well as computer vision and artificial intelligence. Under this paradigm, we seek to find the representation that is *most* *likely* given some image data. Each region of data has a certain probability $P(D|M)$ of being approximated by a particular representation, and each representation occurs with a certain a priori probability $P(M)$. Through Bayes' theorem

$$P(M|D) \ = \ \frac{P(D|M)P(M)}{P(D)}$$

we compute the probability $P(M|D)$ that a representation accounts for a given region. Maximum a Posteriori (MAP) is the principle that we should pick the representation that maximizes this Bayesian likelihood.

$$M_{MAP} \ = \ arg\ max\{P(D|M)P(M)\}$$

This MAP choice is also the minimal encoding of the image when we use an optimal code. Given prior probabilities of various representations, the optimal encoding of each representation uses $-log_2 P(M)$ bits. Similarly, the deviation of a region from a representation can be encoded in $-log_2 P(D|M)$ bits.[16] The MLE estimate for a given image is found by minimizing the combined length of encoding the representation and the residual error.

$$\begin{aligned} M_{MLE} &= arg\ min\{-log_2 P(D|M) - log_2 P(M)\} \\ &= arg\ max\{log_2 P(D|M) + log_2 P(M)\} \\ &= arg\ max\{P(D|M)P(M)\} \\ &= M_{MAP} \end{aligned}$$

Thus, when we have "true priors" to use, and can thus find the Shannon-optimal codes, minimal encoding is equivalent to MAP. When these conventional priors are not obvious, the minimal encoding framework provides us with a good method of approximating them: we simply pick the best encoding we have available. This method is particularly useful in vision problems because it gives us a way to produce estimates using image models that are too complex for calculation of direct priors.

Similarly, simplicity and parsimony have been recognized as essential to notions of representation since the pioneering work of the Gestalt psychologists in the early twentieth century [28]. The *minimum principle* [29] holds that the best representation is the one that accounts for the data with the *simplest* model. The simplest representation is defined as the one with the shortest representation given a set of transformation rules allowed on the data; the search for rules that agreed with human perception was a central focus of their work. Recent researchers in this tradition have used structural rules to define simplicity [30] as well as process models [31]. These transformation rules are equivalent to the prior model probabilities of the Bayesians, or the specification of an encoding by the information-theorists.

2.1 Description by Minimal Encoding

Whether we consider a framework of information, likelihood, or simplicity to judge representations, the minimal encoding paradigm remains applicable; the task is to find the representation that provides the shortest encoding of an image. Using the encoding cost of a representation is a natural bias for an information processing system that has finite resources. Alternatively called the Minimum Description Length principle [15] or Minimum Message Length criteria [32], this framework provides an intuitive and powerful method for recovering non-trivial representations.

Compared to the total number of bits used in a pixel-based representation, there is often little information in natural images; thus in a good representation much compression is possible. Most of an image is determined by the physical (shape) and illumination processes in the scene, together with the optical and stochastic processes of image formation. The parameters of these processes are usually of much lower order than the number of pixels in the image, especially when one considers sequences of images of similar scenes. Recovering the parameters of these processes has long been considered a central task of visual perception.

More generally, we can define portions of an image to be redundant when they contain patterns the visual system has seen before, [1] or can generate from combinations of these previous patterns. Thus if the visual system has encountered many objects with convex surfaces, it will be able to explain new images with convex surfaces (such as the outer surface of my mug) in terms of the previous surfaces it has seen. From a coding standpoint, the system need not store the redundant portions, only the essential information of how to regenerate them. We use

[1] "system" here should be broadly defined; for biological vision systems it also includes a notion of the evolutionary context in which the system developed. E.g. the system is considered to have seen certain stimuli innately.

"descriptions" to represent image regions whose information can be more succinctly expressed based on prior experience. Descriptions are image representations that exploit a prior model to establish salient characteristics (parameters) of the data, and can generate an estimated version of the image to see how well the model held. Each description represents some explicit region of the image, and may overlap with other descriptions.

A visual perception system can use descriptions to take advantage of prior knowledge when it encounters new images. If it finds that a region of an image can be generated by a set of pre-existing descriptions, then it can replace that chunk of the image with pointers to those descriptions and their parameters, shortening the encoding length but preserving the amount of information. For example, a shape description and an illumination description could jointly account for a portion of an intensity image: the stored descriptions for that region would consist of the surface parameters and the direction of the illuminant, together with the function that specifies their interaction.

Biological vision systems deconvolve the image signal into a minimal set of constituent descriptions in an extraordinarily efficient manner; to understand vision we must discover the methods and assumptions needed to perform this task. Finding successful descriptions of this kind are essential to many machine vision problems, ranging from the development of autonomous robots to the exploitation of semantic coding techniques for image transmission. These tasks require models general enough to exploit information in the real-world images likely to be encountered by future applications of machine vision, such as a vehicle navigation system or a wireless HDTV NewsCam.

This thesis investigates the task of recovering a set of descriptions that minimally encodes an image. We do not address directly the question of how certain kinds of descriptions come to be in the descriptive repertoire of a visual system, or whether those descriptions constitute a Shannon-optimal encoding. We will make some conclusions on the kinds of descriptions that will be needed, and focus on the question of what methods can be employed to find the best descriptions given such a repertoire and a particular image.

2.2 Qualitative Descriptions

Recovering a model-based description of an image is a topic that has had much attention in the computer vision and pattern recognition literature. When we restrict the descriptive repertoire to simple models, recovery of the best description is relatively straightforward. With *homogeneous* descriptions, the complexities of the models are all identical, and we can ignore the contribution of the model to the encoding cost when searching for the minimum. We simply find the description that accounts for the image with the minimal residual error. Statistical estimation theory often provides analytic methods for this type of task; for example, we can use linear regression to find the best description among those that model a particular image region with a fixed order polynomial surface. However, when the model becomes complex enough to account for interesting phenomena in images, the notion of "optimal solution" usually becomes ambiguous. These types of descriptions use models of varying complexity and thus we must account for the cost of encoding the model as well.

Various techniques have been proposed to solve such problems, typically applying specific prior knowledge to "regularize" the interpretation of the image[14]. These techniques usually make assumptions about scene structure and image formation over a small neighborhood of the image. However we need to be able to model large scale, 3-D knowledge about object shape, texture, illumination, etc..., in order to obtain descriptions that agree with human perception. Such "information" cannot be practically expressed in low-level, *quantitative* models, without being swamped by the complexities of real imagery.

Alternatively, we can try to use *qualitative* models, which encode high-level shape knowledge using a small number of parameters. The use of such models presents a difficult problem, however, as it is unclear how to relate low-order shape descriptions ("parts") to the detailed, quantitative image information (pixel values) that must be segmented. The problem with recovering descriptions using these models is that there are many possible locations, orientations, and sizes of parts, and parts may occlude each other, so that the solution for a global configuration that minimizes error will be highly non-linear. Statistics over a small window will not adequately solve the problem, since there are critical interactions that can occur at distances up to the width of a part. Indeed, this part-to-pixel problem is an instance of a more general problem that vision systems must face, that of integrating heterogeneous descriptions.

2.3 Integration of Multiple Descriptions

The need to compare the relative utility of descriptions based on heterogeneous models is a core problem of visual representation that is rarely addressed. It is likely that vision systems (or intelligent systems in general) are modular [33], and that they have several functionally distinct methods of analyzing an image, each producing a different type of description. The descriptions used for different tasks, perhaps visualizing 3D shape, recognizing faces or dodging a rapidly approaching projectile, must be somehow integrated into a unified perceptual description.

In machine vision, work on intrinsic images [34] and the primal sketch of Marr [2] were a first step towards modeling the interaction between different types of descriptions. Recently, Markov Random Fields have been used to integrate the descriptions produced by various visual modules.[35] While these methods can succeed at combining local information, they cannot practically model higher-order structure in an image.

Using a description-based minimal encoding framework, we can easily integrate the descriptions from different visual modules. Descriptions can overlap, as long as we explicitly model the function that governs their interaction. The integration task is to find the "information" in the image by finding the set of overlapping descriptions that minimally encodes an image.

We can describe a wide variety of images using this simple model. An integrated description consists of a set H of descriptions,

$$H = \{h_0, h_1, \ldots h_M\}$$

with each description h_m providing good approximation of the image d over some explicit region of the image. The region is denoted by a boolean support function s_m, and the approximated values by $\hat{d}_m$. Assuming the support functions do not overlap, such that there is no more than one description whose support is "1" for a given image point:

$$\sum_{m=0}^{M} s_m[i,j] \leq 1 \qquad for\ all\ i,j$$

then the estimated image is simply the sum of each description's estimated data, masked by its support field:

$$\hat{d}_H[i,j] \;=\; s_0[i,j]\hat{d}_0[i,j] \;+\; s_1[i,j]\hat{d}_1[i,j] + ,\ldots, \;+\; s_M[i,j]\hat{d}_M[i,j]$$

When descriptions do overlap, the integration function f that combines them is no longer just addition:

$$\hat{d}_H[i,j] \;=\; f_H\left(s_0[i,j]\hat{d}_0[i,j],\; s_1[i,j]\hat{d}_1[i,j],\; \ldots, s_M[i,j]\hat{d}_M[i,j]\right)$$

Practically, there will only be a small number of descriptions overlapping at one point, so we can decompose the compound f_H into the sum of descriptions with no overlap, plus the simpler relations $f_{m,n,\ldots}$ between the descriptions $\{h_m, h_n, \ldots\}$ that do overlap.

$$\begin{aligned}\hat{d}_H[i,j] \;=\; & f_{1,4}\left(s_0[i,j]\hat{d}_0[i,j],\; s_4[i,j]\hat{d}_2[i,j]\right)\\ & +\; s_2[i,j]\hat{d}_2[i,j]\\ & +\; f_{3,7,9}\left(s_3[i,j]\hat{d}_3[i,j],\; s_7[i,j]\hat{d}_7[i,j],\; s_9[i,j]\hat{d}_9[i,j]\right)\\ & +\; \ldots\end{aligned}$$

These decomposed functions can be simple and still be useful image models, since the descriptions themselves are quite expressive. A pairwise relation could represent a shape descriptions occluding another shape description $f(x,y) = x$, specular reflection of one description off another $f(x,y) = x + y$, or perhaps the illumination of a shape description by a description of the ambient light $f(x,y) = xy$.[36]

The total error of an integrated description relative to the original image $d[i,j]$ is thus easy to compute.

$$\sum_{i,j}\left(\hat{d}_H[i,j] - d[i,j]\right)^2$$

As we noted above, minimizing this alone will not solve the integration task; trivially, the description set with minimal error will have one description per pixel! We have to take into account the complexity of the models as well as the residuals, using the minimal encoding framework to find the integrated description with the shortest overall encoding length.

The next chapter will discuss the methods we use to estimate individual descriptions which locally provide a minimal encoding. Later, we will address ways to find the subset of these descriptions which constitutes a minimal encoding of the entire image.

Chapter 3

Parallel Robust Estimation

We first address the task of finding a collection of individual descriptions that are candidate members of a minimal integrated description. We will have a set of description models that we can employ to encode the image; our strategy is to test each model over each small neighborhood of the image, and see if we can build a description that offers a substantial savings in encoding the image.

For a given model, there are often simple methods for estimating a single description. However, these methods often fail when objects are partially obscured or occluded, since they treat such corruption as stochastic noise. In order to recover descriptions of complex scenes, we must be able to recover estimates in the face of overlapping objects and processes. Thus we need a framework that allows these estimation methods to model corrupting processes.

Some classes of descriptions are naturally insensitive to certain types of overlap, and we need not worry about those. When a description is in a sense "orthogonal" to the descriptions that are possibly overlapping it, it is possible to directly estimate the description without worrying about the overlap regions. For example, we can estimate the high frequency information in a signal independent of overlapping low frequency information in the image. But many descriptions that we wish to use are not independent of overlapping descriptions, and the parameters we estimate would reflect neither the underlying nor overlapping process accurately. These conventional estimates treat overlap regions as "noise", which skews the estimation process. For these cases we need to develop a new estimation method that is robust in the presense of overlap.

There are many types of overlap processes that occur in real imagery: a shape description

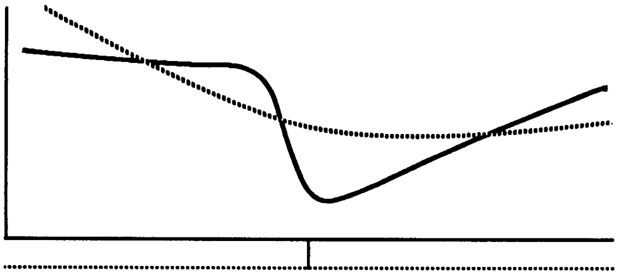

Figure 3-1: Quadratic approximation (dashed curve) of a composite signal (solid curve) consisting of three distinct sections. Estimate is obtained using least squares regression based on support region (dashed bar at bottom) which covers entire signal.

can be occluded by another shape, a description of surface albedo can be added onto a shape description, or a description of ambient light could be combined with a description of surface orientation to account for a rendered intensity image. While all of these would be needed to account for intensity images, so far we have only developed an estimation framework that accounts for the first of these, occlusion, since we have mostly been interested in range imagery.

Occlusion often causes problems with conventional estimates of surface structure. In particular low-order surface models will have poor interpolation at discontinuities where different underlying shapes meet. Figure 3-1 shows a quadratic approximation of a signal which is comprised of three distinct regions; the single-description approximation is poor, since the second-order model can not account for transitions. Yet this is a simple signal! Somehow we should be able to discover that it can be succinctly described by three low-order descriptions.

We have developed an estimation technique that is robust with respect to occlusion: we estimate description parameters only over "supported" points, and exclude from this support those points which we can attribute to a different process. A key aspect of our model is that there are many descriptions being estimated at the same time, perhaps using different models and/or over different neighborhoods of the image. Thus if a particular process is occluding an underlying process, it is likely that there will be some description of the occluding process being estimated. If we can identify this relationship, we can exclude the occluding points from the support of the underlying process, and accurately estimate its true parameters.

Figure 3-2 shows several estimates of the signal used in figure 3-1, using several different initial support regions. If we try a large number of these estimates, a few will be a good coarse

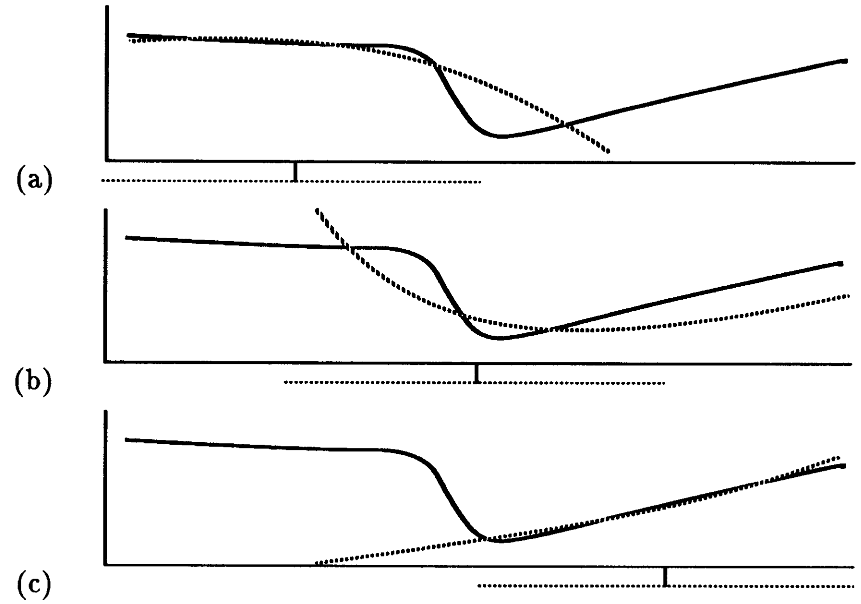

Figure 3-2: (a,b,c) quadratic approximation of previous signal using three different support regions.

approximation to some parts of the signal, as are estimates (a,c) in Figure 3-2. Each estimator can identify which points probably belong to a different process, exclude those contaminating points from its support region, and recover new parameters based on the updated support (figure 3-3.) The essence of our *cooperative robust estimation* framework is that the segmentation of a description is based on the quality of the local fit *and* the quality of all overlapping descriptions. First, we will review some of the traditional robust estimation theory for exclusion of contaminating points based on local criteria, and then discuss our *cooperative* extension to that framework.

3.1 Robust Estimation

The estimation of an underlying process in the presense of interfering processes is one of the central problems of vision and image understanding. Statistical estimators are characterized by two often conflicting measures: their "efficiency", the ability to yield optimal estimates given a certain distributional assumption, and their "robustness", loosely defined as the ability to

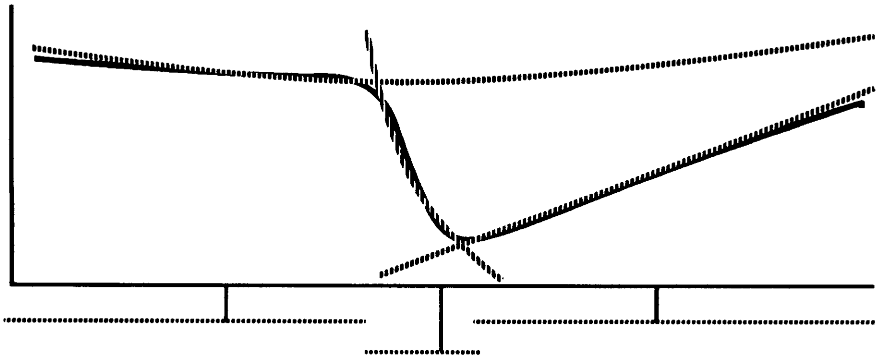

Figure 3-3: Integration of three different estimates which had different initial support regions, after updating support and re-estimating parameter values using the *cooperative robust estimation* technique.

tolerate deviations from the underlying distributional assumption and still return reasonable estimates.

Least squares regression is an example of an estimator that is fully efficient given Gaussian distributed observations, but is sensitive to deviations from this assumption. This estimation method minimizes the sum of squared residuals; thus outliers due to non-Gaussian noise will have an extraordinarily large contribution to the minimization. Statisticians have developed "robust" estimators designed to be less sensitive to the Gaussian assumption.

If a robust estimator can indeed perform well despite outlier noise, it ought to be useful in recovering objects that have contamination due to an occluding process. Recently, several authors have proposed robust mechanisms for image enhancement. One of the most general and popular robust estimation methods, M-estimation, was introduced to the computer vision literature by Besl et. al. [20], for the task of filtering impulse noise while preserving certain kinds of edges.

M-estimators are a class of robust estimators derived from maximum likelihood arguments. Least squares regression is in fact the special case of M-estimation that assumes Gaussian distributed observations. Given a linear system defined by the matrix $\mathbf{A}$,[1] we wish to maximize the probability that the parameters x generated the observed data d. For a given x, the

[1] in our case the columns of the matrix $\mathbf{A}$ are actually 2D images.

estimated data $\hat{d}$ is

$$\hat{d} = \mathbf{A}x$$

and we wish to maximize the probability

$$P = \prod_i exp\left[-\frac{1}{2}\left(d[i] - \hat{d}[i]\right)^2\right]$$

Maximizing P is equivalent to minimizing its logarithm (which is also its optimal encoding length)

$$\sum_i \left(d[i] - \hat{d}[i]\right)^2$$

We can directly solve this by the normal equations since the derivatives with respect to x are linear. However if we generalize to non-Gaussian observations, we now wish to maximize the probability

$$P = \prod_i exp\left(-\rho(d[i] - \hat{d}[i]\right)$$

where ρ is the negative logarithm of the expected density. Defining $\psi = d\rho/dx$, the normal equations which are now

$$0 = \sum_i \psi(d[i] - \hat{d}[i]) \left(\frac{\partial \hat{d}[i]}{\partial x_k}\right) \quad for\ all\ k$$

A Gaussian prior corresponds to an L_2 norm,

$$\psi(r) = r; \quad \rho(r) = r^2$$

which is considered to be "non-robust" in the literature because noise points that do not come from a Gaussian distribution will severly skew the estimation. Some error norms which are considered to be more "robust" are are the L_1 norm

$$\psi(r) = sgn(r); \quad \rho(r) = r$$

blends of L_1 and L_2 norms

$$\psi(r) = \left\{ \begin{array}{cc} r & r < k \\ sgn(r) & r >= k \end{array} \right\}$$

and "redescending norms" that completely exclude outlier points

$$\psi(r) = \left\{ \begin{array}{cc} sin(r) & r < \pi \\ 0 & r >= k \end{array} \right\}$$

In general there are no analytic solutions to these normal equations, and nonlinear equation solvers will often fail since the derivatives of ρ are often discontinuous [25] One solution technique which has been presented for solving the M-estimation problem is Iteratively Reweighted Least Squares (IRLS) [26]. This algorithm approximates an arbitrary norm by the use of an iterative L_2 estimation with a dynamic weighting factor on each point.

We can solve for the L_2 estimate using the inverse of the projection matrix $\mathbf{A}^T\mathbf{A}$

$$x = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T d$$

IRLS adds a weight term on each point to adjust the contribution of that point's residual to the minimization. A diagonal weight matrix $\mathbf{W}$ is computed based on the residuals of each point normalized by a scale factor s,

$$W_{ii} = \frac{\psi\left((d[i] - \hat{d}[i])/s\right)}{(d[i] - \hat{d}[i])/s}$$

and the parameters are obtained by iteratively solving the equation

$$x = (\mathbf{A}^T W \mathbf{A})^{-1}\mathbf{A}^T W d \tag{3.1}$$

At each iteration W is recomputed based on the residual error from the previous step. With a robust error norm, points that deviate excessively from an initial fit will have their significance down-weighted in subsequent iterations.

3.2 Support Update based on Minimal Length Encoding

We can base our choice of a support update function on a minimum encoding length criteria. For a description estimate h_m, with support s_m and estimated data $\hat{d}_m$, [2] we can compute both the number of supported points $N(h_m)$, and the squared residual error over those points, $R(h_m)$:

$$\begin{aligned} N(h_m) &= \sum_{i,j} s_m[i,j] \\ R(h_m) &= \sum_{i,j} s_m[i,j](\hat{d}_m[i,j] - d[i,j])^2 \end{aligned}$$

Using a Minimal Length Encoding paradigm, the description h_m can be considered an *encoding* of the image which saves $S(h_m)$ bits compared to simple pixel-by-pixel description of the image pixel values.

$$S_m(h_m) = k_p N(h_m) - k_r R_m(h_m) - O_m. \tag{3.2}$$

The overhead cost of encoding the estimate, O_m, is the number of bits needed to encode the model parameters and support shape. A simple way to represent the support shape is to chain-code its boundary. We can compute the cost of this code by by computing at each point the number of unsupported neighbors, $b_m[i,j]$, and summing over all supported points.

$$O(h_m) = k_x Rank(x) + k_l \sum_{i,j} s_m[i,j] b_m[i,j]$$

The constants k_p, k_r, k_x, and k_l are the average entropy of raw image pixels, residuals, model parameters, and perimeter chain code, respectively.

In order to locally maximize the encoding savings of a description, therefore, we should use a boolean weight which only includes points which increase that savings. The weights are thus contained in the support field, s_m, which is determined by:

$$s_m[i,j] = \left\{ \begin{array}{ll} 1 & if \;\; (\hat{d}_m[i,j] - d[i,j])^2 \; < \; \frac{k_p - k_l b_m[i,j]}{k_r} \\ 0 & otherwise \end{array} \right\} \tag{3.3}$$

[2] we now switch to 2D notation to make clear the fact we are dealing with images.

But this rule itself is recursive, since b_m depends on s_m; thus we approximate the optimal solution by iteratively solving for s_m, initially assuming b_m is zero. Ideally, after several rounds of support update and parameter re-estimation, the description will have converged to have a large encoding savings.

3.3 Limits on Conventional Robust Estimators

Unfortunately, when faced with even a moderate amount of occlusions, we can rarely be guaranteed that any single robust estimator will converge to a "good" description. Kim et. al. [24] notes the infeasibility of using M-estimation in the presense of transitions between distinct regions, since the level of contamination is beyond what the estimators can handle. The breakdown point of an M-estimator characterizes the limits of robust performance; it is defined as the smallest fraction of contamination which will necessarily affect the value of the estimate. Li [27] has shown M-estimators have breakdown points that are less than $1/(p+1)$, where p is the number of parameters in the regression. Thus even a planar regression fit (with $p = 3$) cannot reliably fit a surface when it becomes more than 25% contaminated with noise or an occluding edge. As an alternative to M-estimation, Kim presents a robust algorithm for planar fits based on least median squares using a Monte-Carlo solution technique. He claims a 49% breakdown point for this method.

But this is unacceptable from our standpoint, since we need an estimator that can find an object that is more than 50% occluded! Since conventional robust estimators treat contaminating processes as unknown noise, these estimators will always be bound by this limit on the amount of contamination they can handle.

3.4 Cooperative Framework

Because of this inability to model contaminating processes, a purely local robust estimator is not well suited to segment occluded regions. However, we can overcome this limitation by estimating several regions at once, extending the traditional framework to explicitly model the occluding processes just as it models the real "signal".

Using an array of parallel estimators, we can cooperatively exclude points based not only on local fit residuals, but also on the residuals of all other fitters that share the point. Such an estimator does not suffer from the critical limits of maximum contamination that affect M-estimators. Given a set of description estimators H,

$$H = \{h_1, h_2, ..., h_N\}$$

support is recomputed in each using a reweighting function that depends not only on the single residual $(\hat{d}_m[i,j] - d[i,j])$, but also on the residuals in other overlapping estimators.

$$\mathcal{R}_{ij} = \{(\hat{d}_0[i,j] - d[i,j])^2,\ (\hat{d}_1[i,j] - d[i,j])^2,\ \dots\ ,\ (\hat{d}_N[i,j] - d[i,j])^2\}$$

$$s_m[i,j] = \psi\left((\hat{d}_m[i,j] - d[i,j])^2,\ \mathcal{R}_{ij}\right)$$

The additional information provided to an estimator by this framework allows the exclusion/segmentation of outliers based not just on their deviation from the prior model, but also on the ability of some other estimator to account for the point in question. It is a reasonable assumption that some other estimator will be attempting to fit an occluding object, which can then bias other estimators towards not including the points of that object in their support.

The cooperative framework outlined above is independent of many of the assumptions of M-estimation, particularly the assumption of linear basis functions. Any estimation or interpolation method that incorporates a boolean support region and has estimated data (e.g. is a description) can be used within this framework. Additionally, there is no constraint that the estimators need be homogeneous; estimators with different basis functions, scales, interpolation strategies, or prior model assumptions can cooperate to describe the signal.

To reduce the computational burden of the support update rule, we restrict our robust estimators to account for occlusion rather than a general overlap function. We can then limit the ψ function to only depend on the minimum residual at a particular image point across all estimators:

$$s_m[i,j] = \psi\left((\hat{d}_m[i,j] - d[i,j])^2,\ min(\mathcal{R}_{ij})\right)$$

We can minimize the total residuals in all estimators by using the "greedy" weight function,

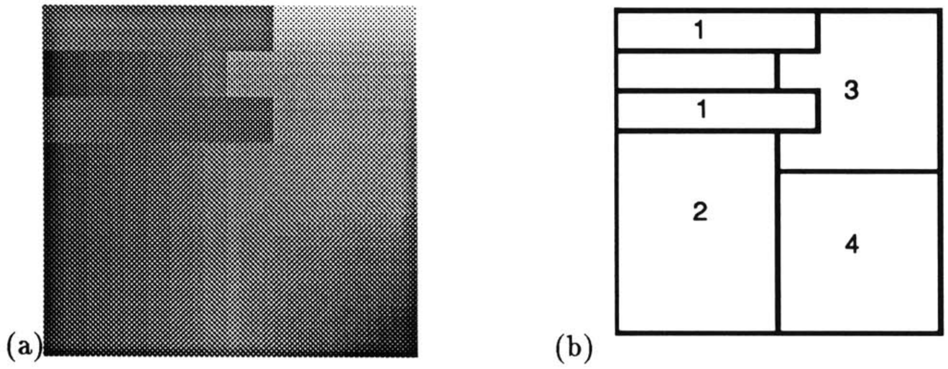

Figure 3-4: (a) A synthetic image with four underlying surfaces. (b) The segmentation of this image into constituent processes/regions. Region (1) is described by $z = y$, region (2) by $z = x$, (3) by $z = y^2$, and (4) by $z = xy$. Thus there is a zeroth order discontinuity between region (1) and (2), a first order discontinuity between (2) and (4), and a second order discontinuity between (3) and (4).

which only allows the estimator with the lowest residual to keep a given point.

$$s_m[i,j] = \left\{ \begin{array}{lll} 1 \;\; if \;\; (\hat{d}_m[i,j] - d[i,j])^2 & < & \frac{k_p - k_l b_m[i,j]}{k_r} \\ \;\;\;\; and \;\; (\hat{d}_m[i,j] - d[i,j])^2 & = & min(\mathcal{R}_{ij}) \\ 0 \;\; otherwise & & \end{array} \right\} \tag{3.4}$$

If there is a one to one correspondence between the set of estimators and the objects in the scene, the greedy rule provides the minimal overall encoding as well as an accurate segmentation of modeled surfaces. Figure 3-4 shows a synthetic image with four different underlying surfaces separated by discontinuities of varying degree. Figure 3-5 shows the ideal description of this image into four different description units, such that each description unit has a support field that exactly covers one process in the image, and recovers the correct parameters of the underlying surface.

Of course, we need to estimate both the surface parameters and support shape of descriptions at the same time. Our initial guesses at support shape will rarely match the real surface segmentation; they can be smaller than the real surface or too large, covering more than one surface. Figure 3-6 shows four description units, each initialized to a small window centered on one of the underlying surfaces. Using surface estimates based on the points in the window,

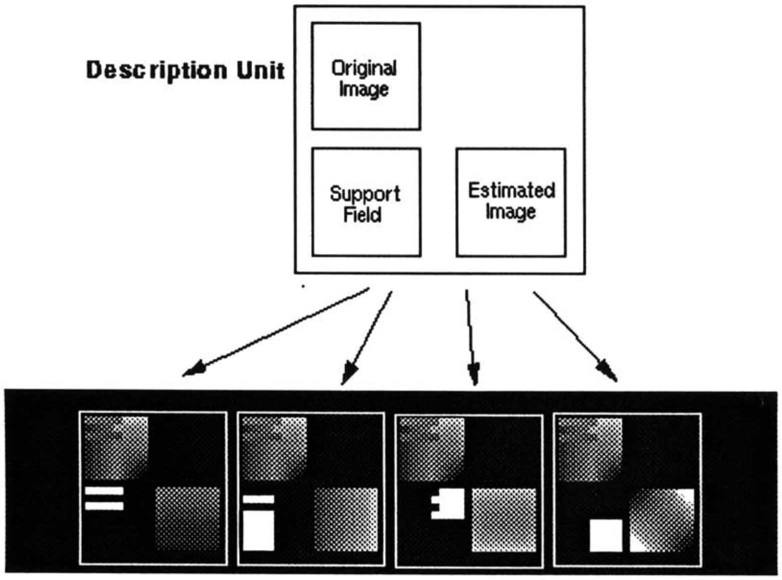

Figure 3-5: The ideal description of the synthetic image in the previous figure with four polynomial description units. Each description has a support shape representing the extent of the various regions, and recovers the surface parameters over that support.

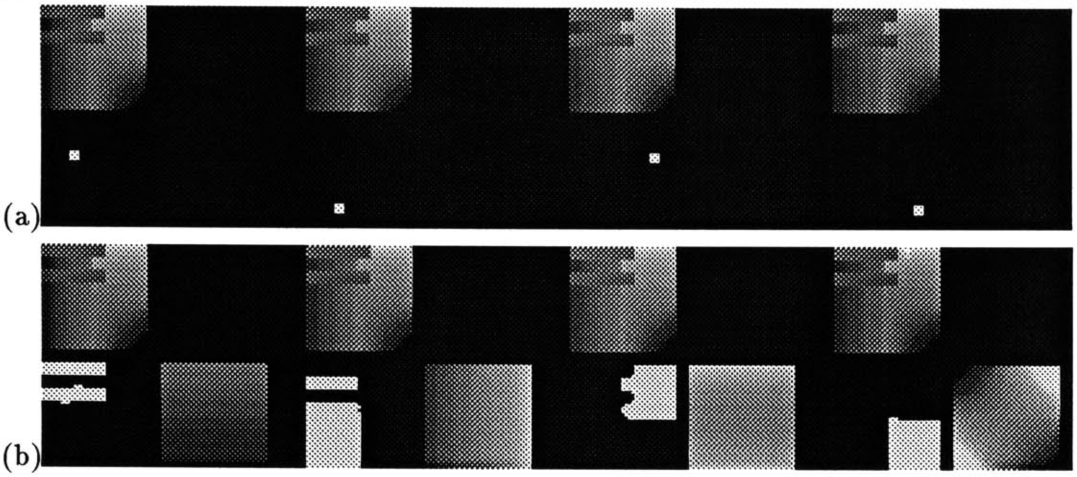

Figure 3-6: Four description units estimating a synthetic image. (a) support shapes set initially to small windows. (b) after estimating parameters and computing residuals based on initial support, support shape is updated using greedy rule. Since the data has little noise, the support easily "grows" to fill the entire actual region based on just a few points.

each estimator was able to "grow" its support region to the correct shape. When a small window of the shape can accurately specify the entire shape, as in this (noise-less) case, small initial supports are very effective. Figure 3-7 shows a similar example, this time using large, overlapping initial supports. At first no estimator makes a correct estimate of the support or parameters. But enough of the support points are correctly segregated after the first estimate so that the support and parameters converge to approximately correct values after several iterations. Indeed, the segmentation for the zeroth and first order discontinuities in the image is quite good, and is respectable the second order discontinuity.

Cooperative estimation will work if, after initial estimation, at least one estimator coarsely estimates each object in the scene. The only way we can make this likely to happen is if there are many estimators with different models and initial support shapes centered throughout the image. But this also means that several estimators will be fitting each object, which violates the one-to-one assumption of the greedy rule. Since the greedy rule does not take into account the description overheads, several partial descriptions of an object can have lower residuals and thus be erroneously considered a better encoding than a description of the entire object. Figure 3-8 shows nine estimators with small initial windows; with the strict greedy rule, support over-segmentation results even when the parameters are almost correct.

The following chapter will discuss the encoding-based optimization methods we employ to take into account the full description overhead and eliminate the redundant elements of a hypothesis set. But we still must generate some initial hypotheses, realizing that we only have an estimate $\hat{\mathcal{R}}$ of the ideal description set residuals. Thus there is some uncertainty around the minimum values, and so we only require points in a local estimate to be within some threshold of the minimum residual:

$$s_m[i,j] = \left\{ \begin{array}{ll} 1 & if \;\; (\hat{d}_m[i,j] - d[i,j])^2 \; < \; \frac{k_p - k_l b_m[i,j]}{k_r} \\ & and \;\; (\hat{d}_m[i,j] - d[i,j])^2 \; <= \; min(\hat{\mathcal{R}}_{ij}) \; + \; \theta \\ 0 & otherwise \end{array} \right\} \qquad (3.5)$$

Initially, when we have far more estimators than objects in the scene, θ is very large, and we will compute a set of description hypotheses whose members share many overlapping points. A moderate threshold alleviates the oversegmentation of the previous example, as shown in

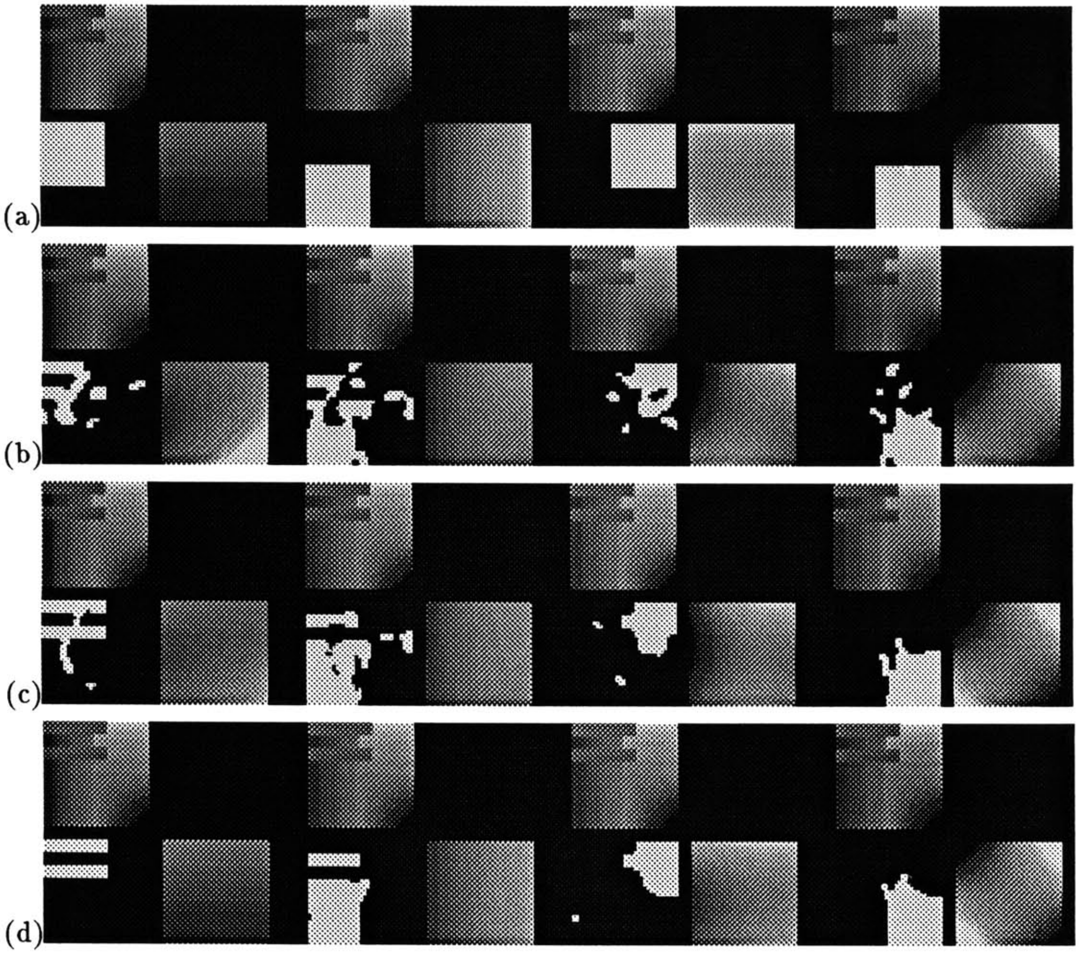

Figure 3-7: Four description units estimating a synthetic image. (a) support shapes set initially to overlapping windows that cover more than one process. (b) the parameter and support estimates are corrupted by these "outlier" points. (c,d) iteratively refined estimates achieve a good segmentation along the zeroth and first order discontinuities.

Figure 3-9. After redundant estimators are eliminated by the encoding length optimization, cooperative estimation can be run with θ close to zero.

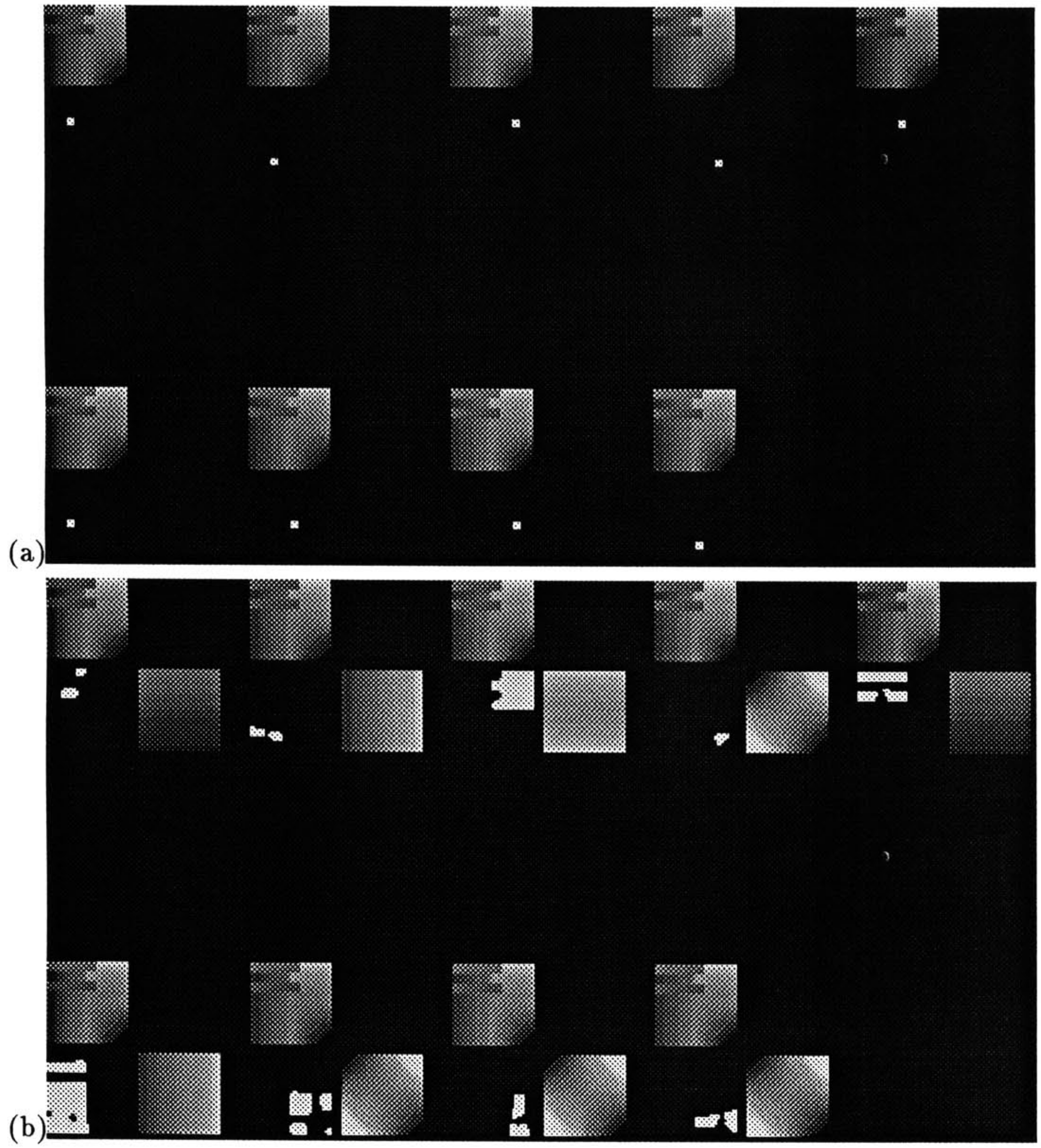

Figure 3-8: Nine description units estimating a synthetic image with four underlying processes. (a) Description units with support shapes initialized to be small windows at evenly spaced points in image. (b) description units after surface estimation and support update based on strict greedy rule. Since there are too many estimators, the support shapes are oversegmented.

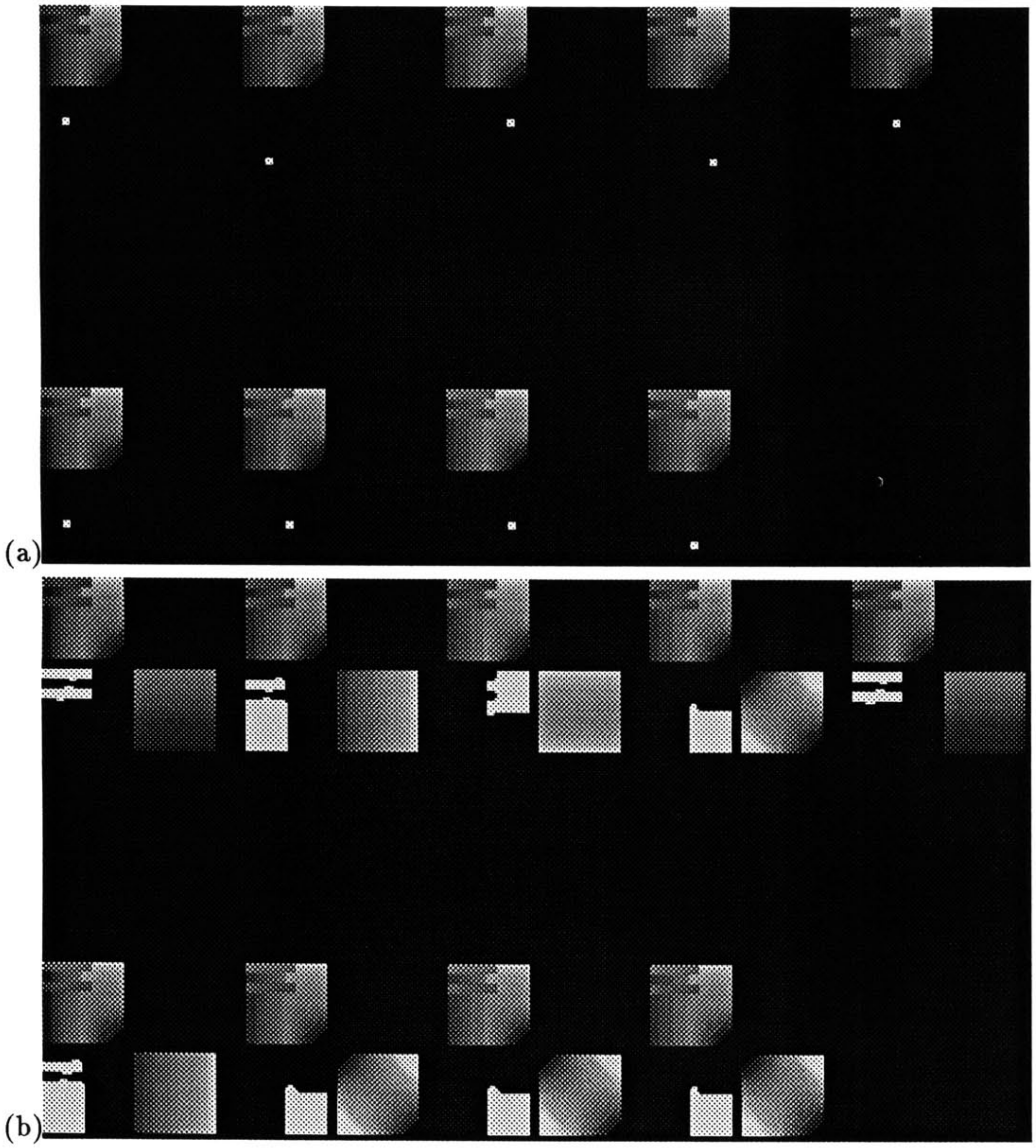

Figure 3-9: (a) description units with initial support shapes. (b) after surface estimation and support update using thresholded greedy rule.

Chapter 4

Network Optimization

We can refine our estimate of the description set by finding the subset which provides the best overall account of the original image, and eliminating the remaining redundant descriptions. We base this choice on the encoding length of a subset, and use a modified Hopfield-Tank network to find the subset with minimal encoding length.

If the effects of the estimates on image encoding are independent from each other, we could simply evaluate the estimates according to the savings they individually afford in encoding the image, picking the single best estimate at each point or fixed region of the image. However, our estimates are not constrained to cover fixed regions, and the perceived likelihood of a description as an explanation of the data is rarely independent of how it interacts with other descriptions. A description whose support is a proper subset of another description will usually be discarded, even though it could pass an arbitrary local threshold. To measure the cost of encoding an image with a *set* of hypotheses, we must be able to account for interactions between the various members of the hypothesis set.

4.1 Accounting for description overlap

When two estimates h_m and h_n overlap, their joint contribution to the overall savings will be different than the sum of their individual savings. The sum $S(h_m) + S(h_n)$ is counting the points in the intersection twice; we can count the number of affected points by summing the product of the two boolean support arrays.

$$N(h_m \cap h_n) = \sum_{i,j} s_m[i,j] s_n[i,j]$$

With occluded estimates, the joint savings is obtained by simply subtracting the excess savings from the sum of individual savings.

$$S_{n-occluded}(h_m \cup h_n) = S(h_m) + S(h_n) - k_p N(h_m \cap h_n) + k_r R_n(h_m \cup h_n)$$

where $R_n(h_m \cup h_n)$ is the sum of the residuals of h_n over the region of intersection with h_m.

However, with estimates that are transparently overlapping, [1] the joint savings is more complicated, perhaps even greater than $S(h_m) + S(h_n)$! In general, we correct for duplicated intersection in the individual savings by subtracting out the savings for the points counted twice, and adding back a residual correction term RC_f which depends on the function f which generates the intersection.

$$S_f(h_m \cup h_n) = S(h_m) + S(h_n) - k_p N(h_m \cap h_n) + K_r RC_f(h_m \cap h_n)$$

The residual correction term for a pair of estimates h_m,h_n depends on what overlap process is generating the region covered by $h_m \cap h_n$. Given an overlap function $f(x,y)$ that defines how two overlapping points are combined, the correction at each point is to add both single-estimate residual errors back to the global savings, and subtract the new joint residual error.

$$rc_f(\hat{d}_m, \hat{d}_n, d) = (\hat{d}_m - d)^2 + (\hat{d}_n - d)^2 - (f(\hat{d}_m, \hat{d}_n) - d)^2$$

We sum this over the region of intersection to obtain the correction term $RC_f(h_m \cap h_n)$:

$$RC_f(h_m \cap h_n) = \sum_{i,j} s_m[i,j]\ s_n[i,j]\ rc_f(\hat{d}_m[i,j], \hat{d}_n[i,j], d[i,j])$$

If h_n is being occluded by h_m, $f(x,y) = x$ then this simplifies to just adding back the residuals

[1] while transparent overlap is not yet accommodated by the current cooperative estimation framework, we have kept the generalization in this chapter.

of h_n over the intersection region:

$$\begin{aligned} RC_{n-occluded}(h_m \cap h_n) &= \textstyle\sum_{i,j} s_m[i,j]s_n[i,j](\hat{d}_m[i,j] - d[i,j])^2 \\ &= R_n(h_m \cap h_n) \end{aligned}$$

Additive and multiplicative functions can be used as rough models to describe "intrinsic" specular and diffuse reflection, as well as accounting for transparent objects.[36] Given a collection of such overlap functions, we can evaluate each of them over the intersection region, and choose the one that has the lowest joint residual. We define $RC_*(h_m \cap h_n)$ to be the corresponding residual correction term, and label the estimate pair accordingly.

Ignoring points in the original image where three or more hypotheses overlap, we can construct matrices based on the intersection terms N and RC_* to compute the savings for a subset of H. We represent an arbitrary subset of H by constructing a bit vector $\vec{x}$ of length M, whose i-th element is 1 if h_i is in the subset. The savings generated by encoding the image data using this subset is

$$S(\vec{x}) = k_p \vec{x}\mathbf{A}\vec{x}^T - k_r \vec{x}\mathbf{E}\vec{x}^T - \vec{x}\mathbf{O}\vec{x}^T. \tag{4.1}$$

with matrices $\mathbf{A}, \mathbf{E}$ and $\mathbf{O}$:

$$a_{ij} = \left\{ \begin{array}{cc} N(h_i) & i = j \\ -\frac{1}{2}N(h_i \cap h_j) & i \neq j \end{array} \right\}$$

$$e_{ij} = \left\{ \begin{array}{cc} R(h_i) & i = j \\ -\frac{1}{2}RC_*(h_i \cap h_j) & i \neq j \end{array} \right\}$$

$$o_{ij} = \left\{ \begin{array}{cc} O_i & i = j \\ 0 & i \neq j \end{array} \right\}$$

Equation (4.1) can be extended to include overlaps between three or more descriptions by adding additional terms that express higher-order overlaps. However these higher-order overlap terms are expensive to calculate; moreover such high-order overlaps seem to be infrequent in real imagery. Instead, we have chosen therefore to assume that the *final* solution will contain a negligible number of image points covered by three or more hypotheses. Note that we are *not* assuming that this is true for the entire set H, where such high-order overlaps will be common.

4.2 Hopfield-Tank Optimization

The form of equation (4.1) allows us to construct a Hopfield-Tank network to maximize $S(\vec{x})$[23, 37]. Each description hypothesis h_m has a corresponding "neural" unit with floating-point activity level u_m. The activities are updated according to the differential equations

$$C\frac{du_m}{dt} = \sum_{n}^{M} T_{mn}\, f(u_m) + I_m$$

using a zero-offset sigmoid update function f, and a constant capacitance C for each unit. (We assume infinite resistance in our neural model and this omit the leakage term.) The input activity I_m for a unit is the encoding savings of that description, and the inhibitory weight T_{mn} between two units is the correction term for their intersection.

$$I_m = k_p N(h_m) - k_r R(h_m) - O(h_m) \tag{4.2}$$

$$T_{mn} = k_r RC_*(h_m \cap h_n) - k_p N(h_m \cap h_n)$$

This network will converge to a local maximum of $S(\vec{x})$, with positive u_m reflecting the membership of h_m in the solution set. The convergence of this network to a *global* maximum of $S(\vec{x})$, is guaranteed only when the equivalent solution matrix $\mathbf{Q}$ is positive or negative definite.

$$\mathbf{Q} = k_p\mathbf{A} - k_r\mathbf{E} - \mathbf{O} \tag{4.3}$$

Unfortunately this is not typically the case for our integrated description task. While the network is guaranteed to converge (the inhibitory weights are symmetric,) it will often do so at a local minima that is far from the global minima. We use a *continuation method* to avoid these local minima, first choosing a related problem that can be solved, and then iteratively solving a series of problems that are progressively closer to the original problem.

In the problem at hand, $\mathbf{Q}$ is easily solved when the diagonal values of $\mathbf{O}$ are sufficiently large, since $\mathbf{Q}$ is then diagonally dominant and negative definite. We can obtain an initial solution by solving with a constant K added to each O_m, or equivalently subtracting K from the input activities of our network. With a sufficiently large K, $\mathbf{Q}$ will be diagonally dominant, and thus the network will converge to a good global solution (for that value of K.) [37]

$$C\frac{du_m}{dt} = \sum_{n}^{M} T_{mn}\, f(u_m) + I_m - K$$

We allow the net to stabilize for a particular value of K, and reset units which have negative activity. Units which have achieved positive activity are left unchanged, and the net is run with a lower value of K. This continues for several iterations until an acceptable solution is found or K is zero. Because K effectively acts as a threshold on the allowable cost of adding a new hypothesis to our description set, the effect of this continuation technique is to solve for the largest, most prominent structures first, and then to progressively add in smaller and smaller detail until the entire figure is well-approximated.

Chapter 5

Part Segmentation

We have applied the cooperative estimation and optimization methods described in the previous chapters to develop a system which segments range images into their constituent surfaces. *Part* models are the basis for our descriptions, since many interesting objects often consist of approximately rigid chunks connected by hinges or other joints. A segmentation of such an articulated object into its constituent parts can be both a perceptually efficient representation as well as an important construct for modeling and thinking about the "real world" [1, 2, 3, 4, 5, 6, 8]. Parts are typically "blob-like" volumetric primitives; however, since range images only have $2\frac{1}{2}$D scene information, we use a surface-based approximation. A prototype of our system has been implemented on a Connection Machine 2, and tested on both real and synthetic range imagery. The present implementation uses low order polynomial patches within a bounding ellipse to describe part surfaces. Since the only intersection we expect between objects in a range image is occlusion, that is the only overlap function allowed.

5.1 Architecture

Our architecture is centered around a network of *description units,* each of which attempts to model a region of the image, and has links to all other overlapping description units (figure 5-1.) The units are densely packed over the image, with several at each image location. [1]

[1] In the most recent implementation, we use a multi-resolution representation of the original scene, so that some of the units have a coarser image of the data, but cover a wider field of view.

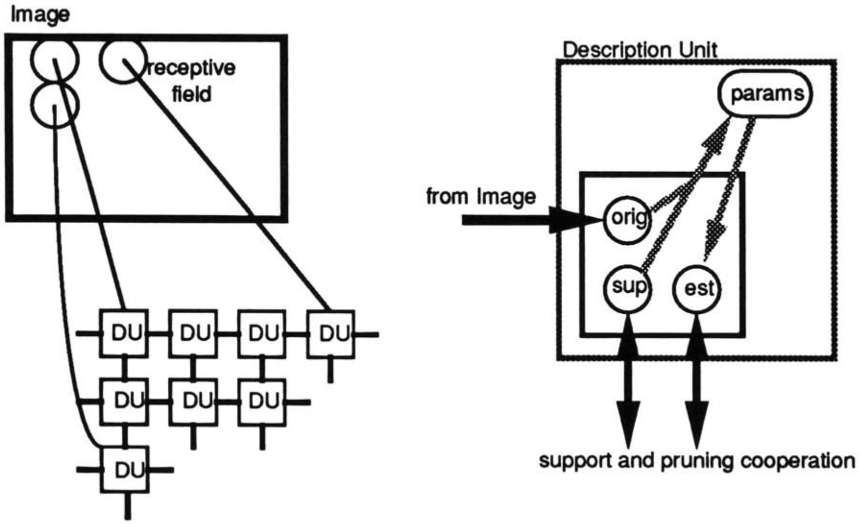

Figure 5-1: Network of Description Units

Each description unit attempts to recover a description which offers the (locally) maximum encoding savings. Initially, there are up to eight descriptions units for every point in the image, each using a different initial support shape. Since we are using a polynomial surface model, we can use linear combinations of a set of basis functions to approximate the original data.

An iterative least squares estimation finds the surface and support shape which minimizes squared residuals and perimeter encoding cost. Parameters are estimated based only on the supported points; at each iteration the support field is updated to contain the points that are well described by the estimate, in terms of both the residual error and perimeter encoding cost. Since we know that there will be far more description estimates in this initial set that real objects in the scene, cooperative estimation is not used at this stage (e.g. θ is infinite.)

Figure 5-2 shows a subset of description units that remained after initial estimation on a range image of a model man. While some of these are fairly good estimates of a “head” part, the local robust estimation has had a difficult time segmenting points that belong to other parts. We need to use cooperative robust estimation to achieve a good segmentation, but first we must prune the redundant estimates from our hypotheses set.

An encoding-length based optimization is used to decide which elements of the hypotheses set to retain. A local pruning algorithm eliminates descriptions which are substantially oc-

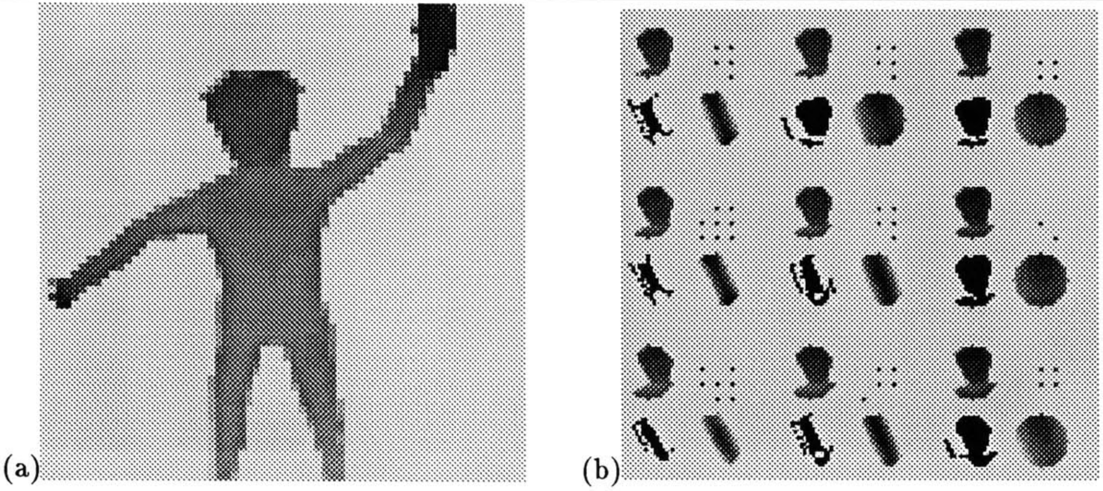

Figure 5-2: (a) a range image of a CAD man (b) 9 description units centered on the region of the range image containing the man's head. Each description unit shows three subfields: the original data (upper left), support region (lower left), estimated data (lower right.) (Dots in the upper right indicate certain status information.)

cluded by an adjacent estimate. In Figure 5-2, the estimate in the center right would probably "prune" the others using this local rule. This heuristic is only valid when a pruned part is likely to have been actually representing the same surface, hence the restriction to only prune adjacent estimates. To further prune estimates we need a more global criteria that accounts for the interaction between all overlapping descriptions. The Hopfield-Tank network presented in Chapter 3 will do just that, converging to the subset of hypotheses which constitutes a local minima of overall encoding. We find a subset with a globally short encoding by using this net in a coarse-to-fine manner, using a scale-space continuation method to avoid local minima. [2]

Since the number of description hypotheses have been significantly reduced after running

[2] In the multiresolution implementation, links between units in the same level of the pyramid are computed in parallel, but the links that cross levels are computed only as needed as the coarse-to-fine method progresses. The network is run first only for the coarsest level of the pyramid, using a value of K such that only the top level descriptions cover enough of the image to exceed the savings threshold, and converges such that the units with positive activity represent a set of descriptions with a short overall encoding. The coarse-level descriptions that did not make this set are discarded from the net; for those that did, inter-level links are computed with all other descriptions in lower levels. We then repeat this process for each lower level in the pyramid, lowering the value of K accordingly.

In the process of computing the inter-level links from a description to finer scale descriptions, we compute a full-size, multi-resolution instance of the description. The receptive field of these full-resolution descriptions are no longer limited to a small neighborhood of the image. The support for these estimates can potentially cover every point in the image, all of which would then be used in estimating the surface parameters. This allows correlation and common structures to be found when they are far apart and at fine scale.

the optimization net, we can employ cooperative robust estimation to refine the support fields. We lower the support threshold θ, and compute new support fields and parameter values for each description. Using these revised descriptions, we update the network links to reflect the new intersection regions, and re-run the network to further prune "unnecessary" hypotheses. A few more iterations of estimation and pruning occur until either the overall description length is satisfactory, or will decrease no further.

5.2 Segmentation Results

We have implemented a prototype of this architecture on a Connection Machine 2 with 16K single-bit SIMD processors and 64K bits of memory per processor. Given an input image with up to 128 by 128 total pixels, we can consider up to 8 initial structure estimates at each point, each estimate having a window of up to 16 by 16 pixels. Up to 2,048 of these estimates can participate in the initial multi-resolution Hopfield-Tank network, and up to 128 descriptions can be considered in the full-resolution estimation/ network optimization. This system was implemented entirely in *lisp, and the typical run time for the following examples was on the order of 4 minutes.

5.2.1 Range image segmentations

We ran our descriptive network on a variety of range images, both natural and synthetic: Figure 5-3a shows a range image of a collection of blocks, [3] for which a minimal part-based description set was computed. The support shapes for the descriptions in this set are shown pseudo-colored in Figure 5-3b. The estimated data for these descriptions are shown in Figure 5-3c. Figure 5-4a shows a range image of a cylinder and blocks; Figures 5-4b,c show the segmentation and estimated data for the recovered minimal descriptions. Figure 5-5a shows a complex synthetic scene with a non-constant background. The estimated data in the recovered descriptions is shown in Figure 5-5b. The pseudo-color segmentation is shown in Figure 5-5c

[3] Range images in figures 5-3 and 5-4 courtesy of MSU Pattern Recognition and Image Processing Lab

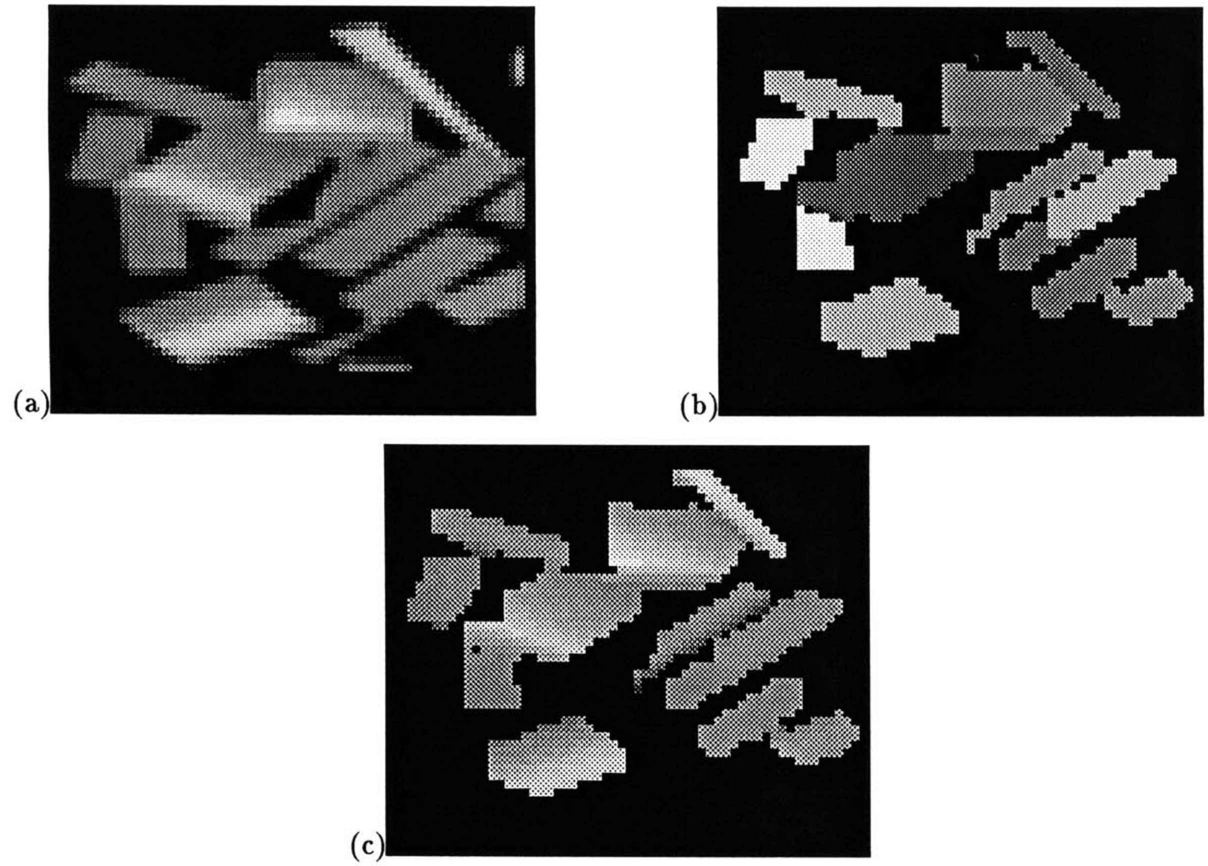

Figure 5-3: (a) range image of assorted blocks. (b) recovered segmentation. (c) combined description estimates.

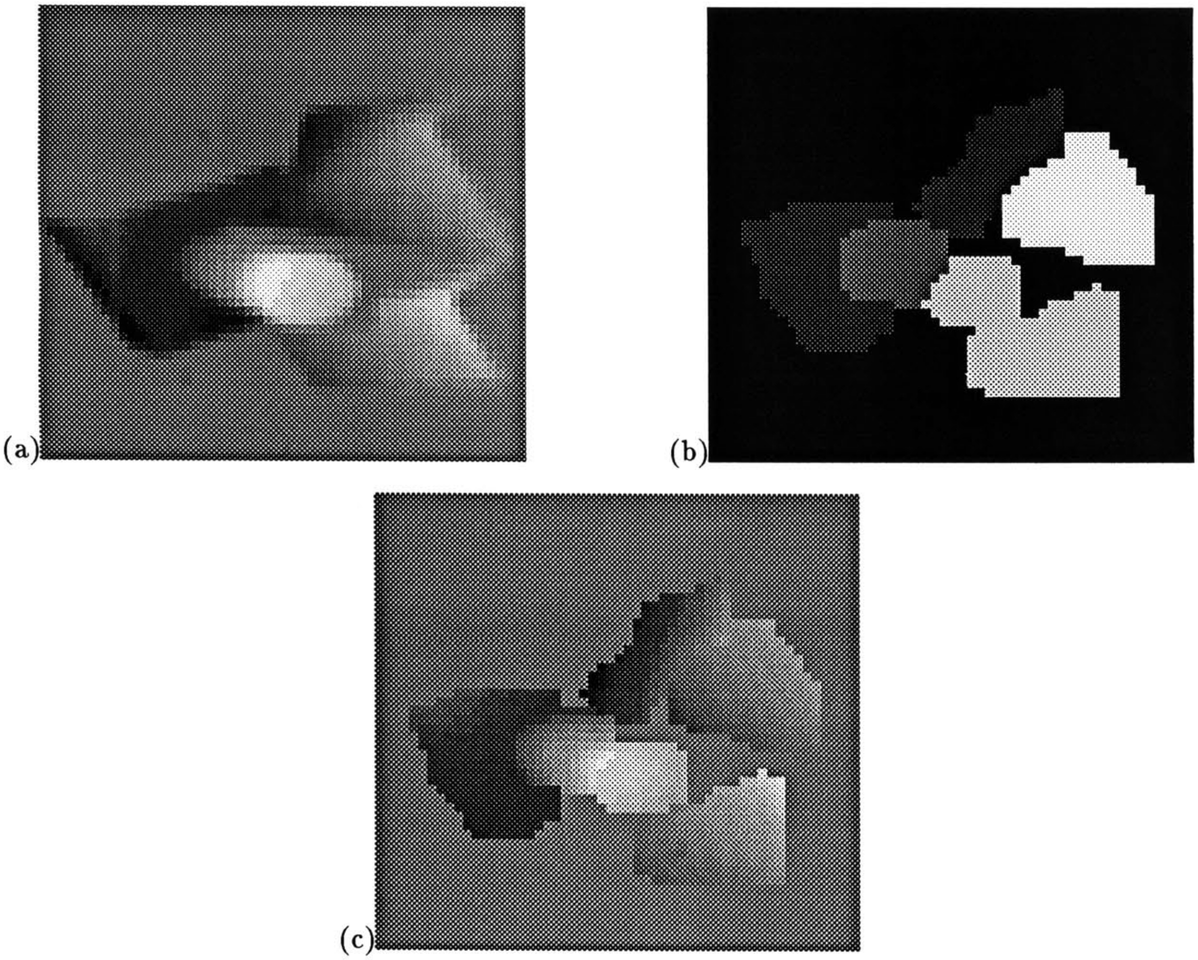

Figure 5-4: (a) range image of a cylinder and blocks. (b) recovered segmentation. (c) combined description estimates.

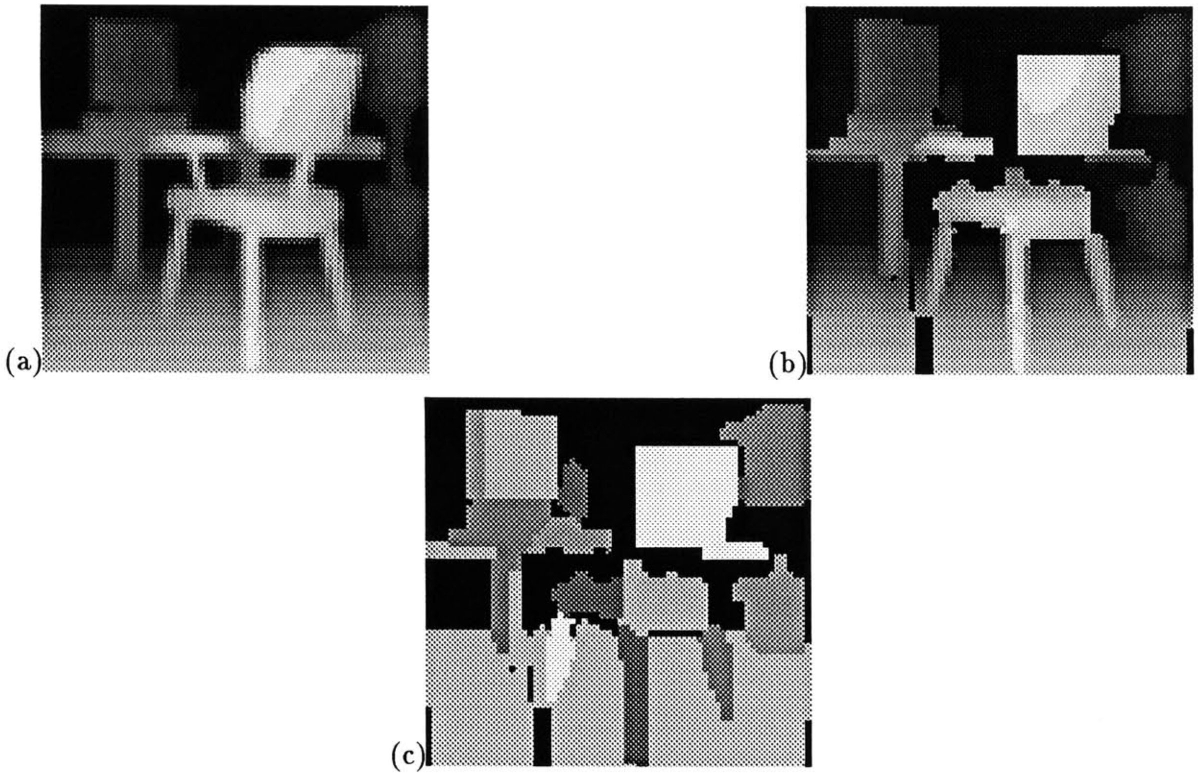

Figure 5-5: (a) range image of scene. (b) combined description estimates. (c) recovered segmentation.

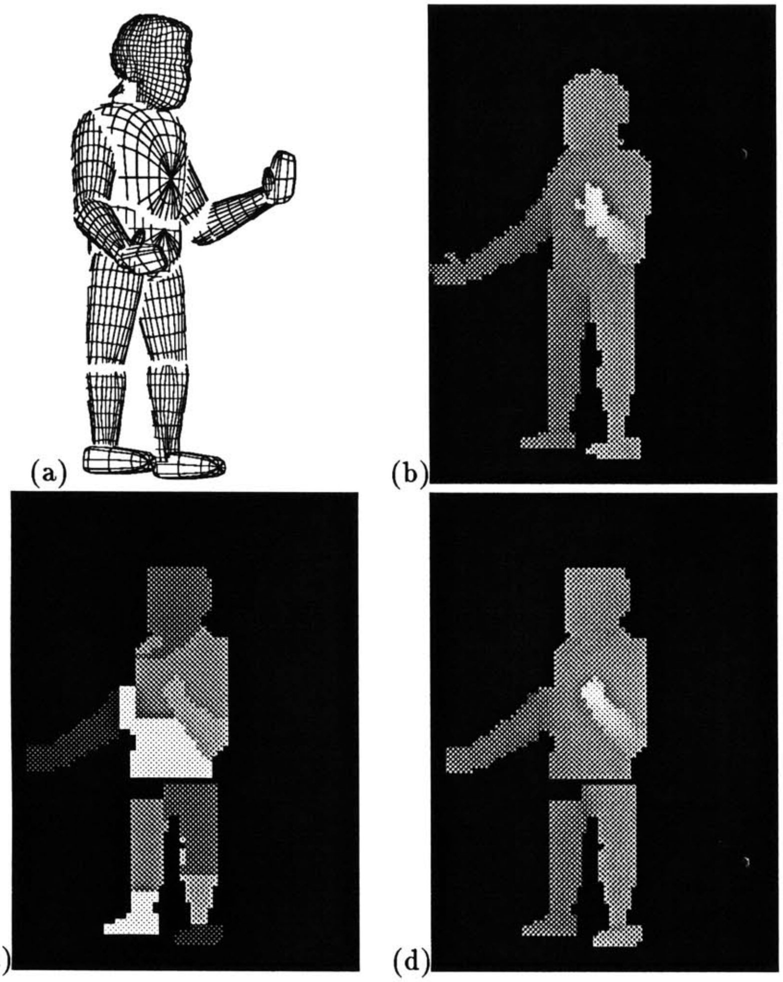

Figure 5-6: (a) CAD model of a man. (b) rendered z-buffer. (c) recovered segmentation. (d) combined description estimates.

Figure 5-6a shows a CAD model of a human figure, which was used to render the “z-buffer” range image in figure 5-6b. This was used as input to our minimal description algorithm, resulting in the segmentation shown in figure 5-6c and the estimated data image in figure 5-6d.

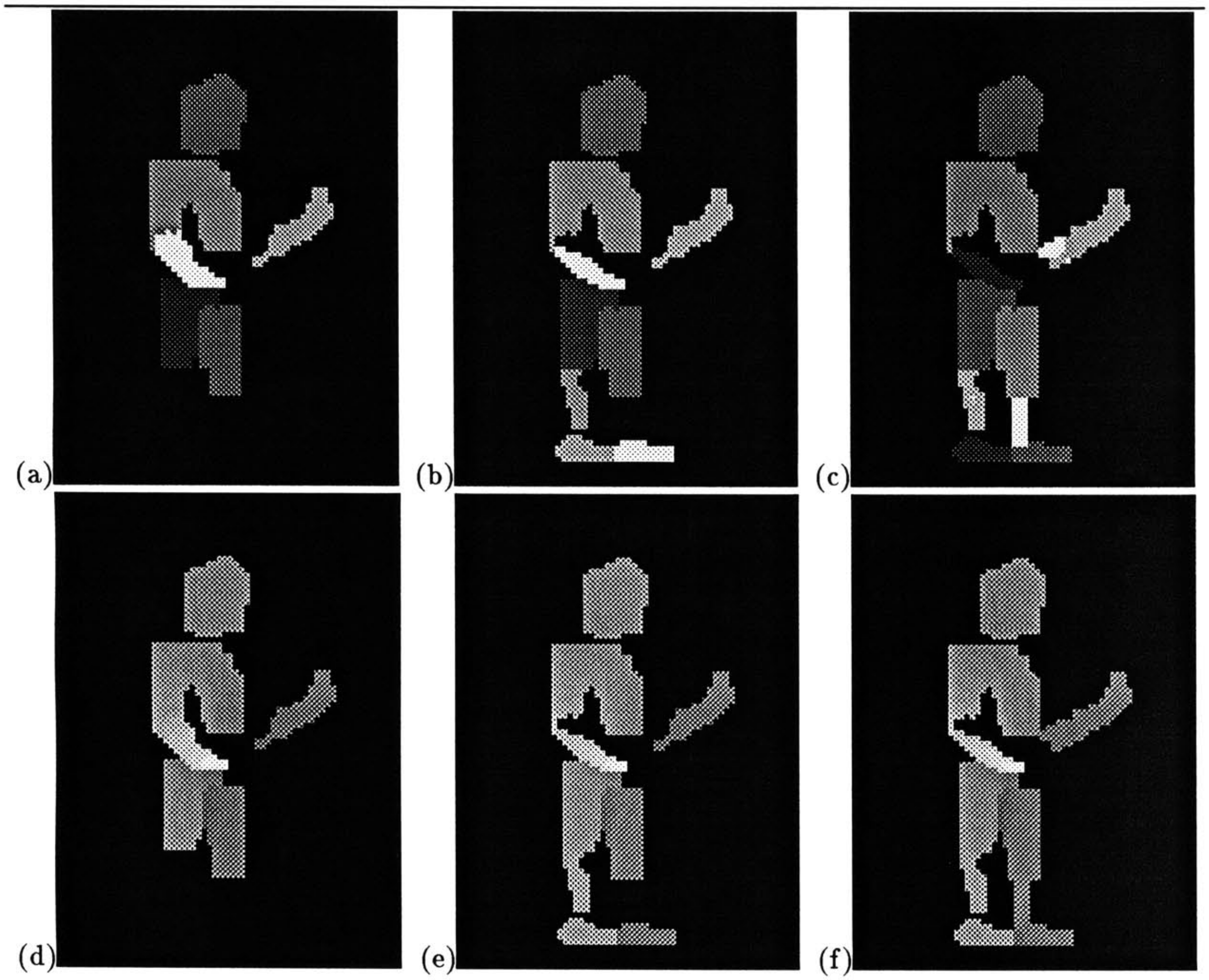

Figure 5-7: (a-c) segmentation after network convergence at various values of K. (e-f) combined description estimates.

5.2.2 Segmentation Stability

We examined the evolution of the segmentation during various stages of the continuation method used by the optimization net, taking snapshots of the active descriptions after convergence for a particular value of K. Figure 5-7 shows the segmentation and estimated descriptions for three of these snapshots; large, prominent parts are recovered with big values of K and remain for the rest of the solution, decreasing values of K fill in the smaller, detail parts.

The segmentation was also stable across multiple views of the CAD man model. Figure 5-8 shows several z-buffers rendered from different camera viewpoints, and the recovered segmentation.

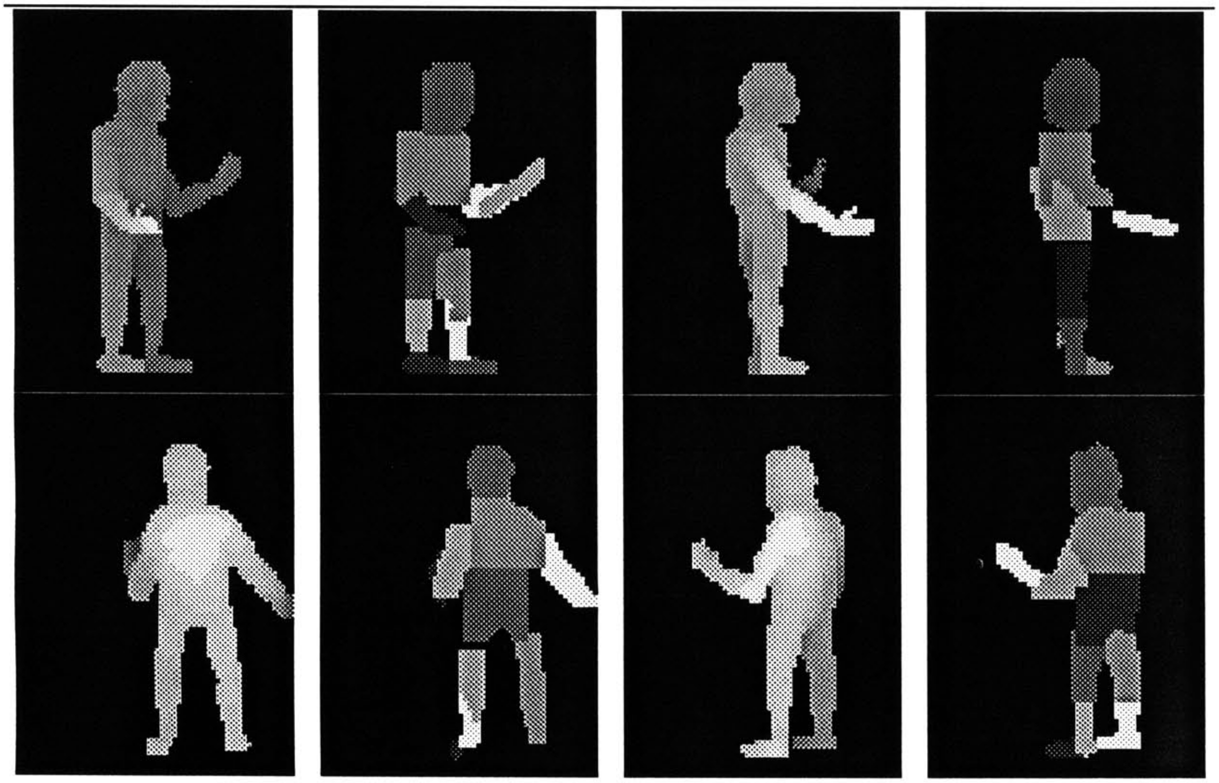

Figure 5-8: Range image of CAD model and recovered segmentation using several different camera views.

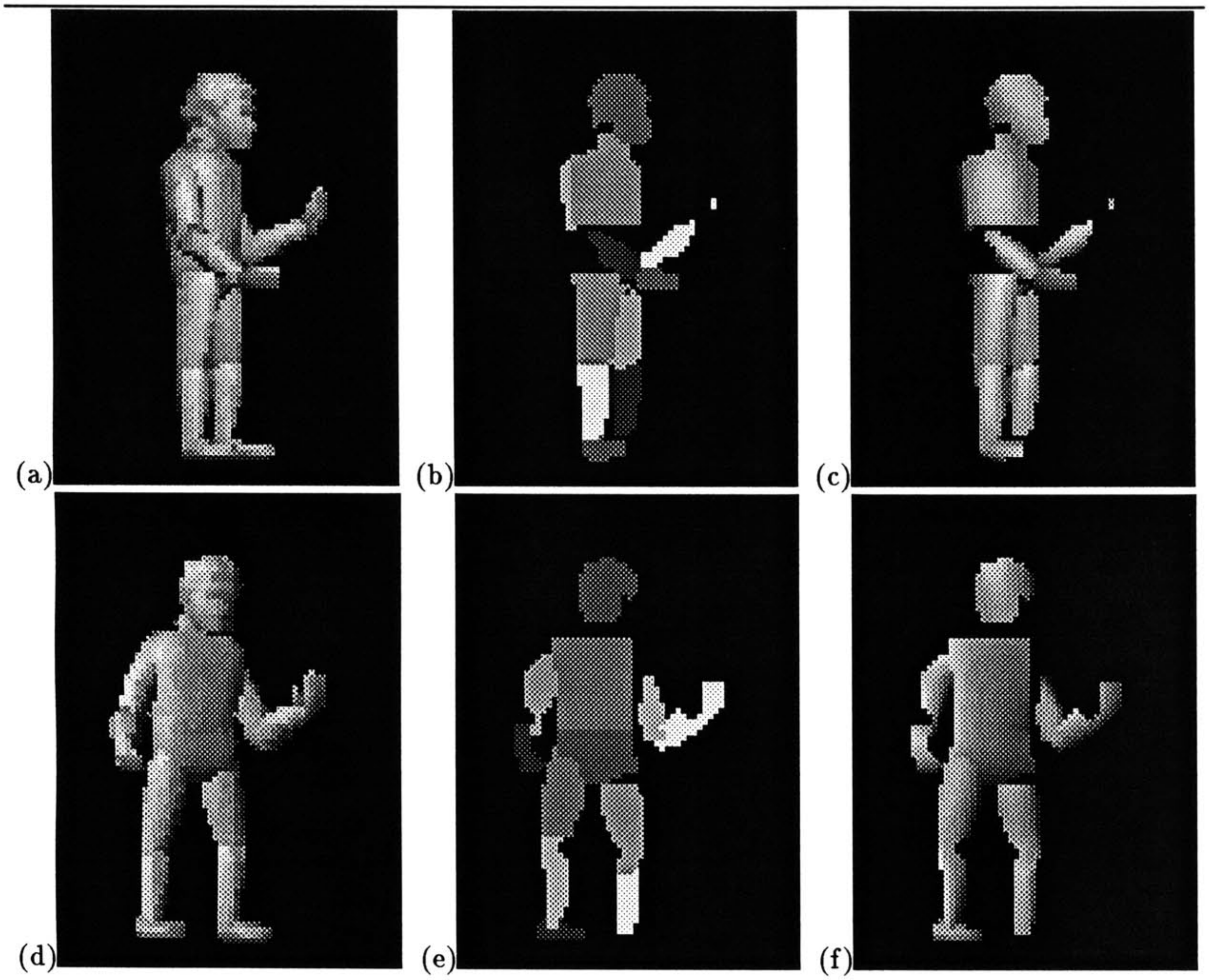

Figure 5-9: (a,d) rendered intensity images of CAD man. (b,e) recovered description segmentations (c,f) combined description estimates

5.2.3 Intensity image segmentation

While we intended our part surface model to be used for range imagery, it can also can be used to describe certain (limited) intensity images! A low order polynomial model is an acceptable description of constant-albedo shapes when illuminated from a single light source without shadows. Thus we tested our system on a Giroud-rendered image of our model man, as shown in figure 5-9. We used the same parameters as in the range segmentation, except for a higher signal to noise ratio. The segmentation was quite good, and in qualitative agreement with the range image result.

Chapter 6

Conclusions

A new paradigm for segmentation has been proposed, in which an image is parsed into the most concise set of descriptions possible. Constituent descriptions use models which reconstruct some portion of the image using a small set of parameters, and may overlap with other descriptions to describe a region of the image where several processes are occurring. By minimizing the overall encoding length of an integrated description, we find the representation that is not only the "simplest", but also the most probable in an information-theoretic sense.

Directly recovering an integrated description of an image is difficult, due to the codependence between estimating the parameters of a description and determining its support. Both parameters and support can be estimated by maximizing the encoding savings of a given description, but only when we are close to a final solution and know the correct number of estimators. But to get close to a solution, we must make many guesses at possible image structure using the description models we have on hand, which means there will be many redundant descriptions in our initial set. Thus we can only approach the solution slowly, first recovering a large set of coarse estimates, then pruning the descriptions which are redundant, and finally refining the parameter and support estimates in the remaining descriptions.

We demonstrated an implementation of a framework which used "part"-based descriptions to decompose images of articulated objects into their constituent shapes. A cooperative robust estimation technique provides occlusion-insensitive initial hypotheses of local shape. We then construct a Hopfield-Tank network based on the encoding savings each hypothesis affords, and on the encoding savings in the intersection regions between hypotheses. Using a scale-space continuation method to avoid local minima by finding the most prominent parts first, this net finds the subset of hypotheses which minimally encodes the image. Initial experiments on scenes with significant part structure yielded successful results. Segmentations of real and synthetic range images were perceptually salient and stable across viewpoint; rendered intensity images that fit our model were also easily segmented by this system.

Bibliography

[1] Binford, T. O., (1971) Visual perception by computer, *Proceeding of the IEEE Conference on Systems and Control*, Miami, December.

[2] Marr, D. and Nishihara, K., (1978) Representation and recognition of the spatial organization of three-dimensional shapes, *Proceedings of the Royal Society - London B*, 200:269-94

[3] Hoffman, D., and Richards, W., (1985) Parts of recognition, In *From Pixels to Predicates,* Pentland, A. (Ed.) Norwood, N.J.: Ablex Publishing Co.

[4] Pentland, A., (1986) Perceptual Organization and the Representation of Natural Form, *AI Journal*, Vol. 28, No. 2, pp. 1-38.

[5] Beiderman, I., (1985) Human image understanding: recent research and a theory, *Computer Vision, Graphics and Image Processing,* Vol 32, No. 1, pp. 29-73.

[6] Konderink, Jan J., and van Doorn, Andrea J., (1982) The shape of smooth objects and the way contours end, *Perception, 11*, pp. 129-137

[7] Shannon, C.E. (1948) *Bell System Tech. Jour.* Vol 27, 379-423

[8] Leyton, M. (1984) Perceptual organization as nested control. *Biological Cybernetics 51*, 141-153.

[9] Pentland, A., (1990) Automatic recovery of deformable part models, *Intl. J. Computer Vision* Vol 4, p 107-126.

[10] Burt, P. J., (1983) Fast Algorithms for Estimating Local Image Properties, *Computer Vision, Graphics, and Image Processing,* Vol. 21, pp. 368-382.

[11] Hopfield, J. J., and Tank, D. W., (1985) Neural computation of decisions in Optimization Problems, *Biological Cybernetics,* Vol. 52, pp. 141-152.

[12] Ballard,D.H., Hinton, G.E., and Sejnowski, T.J., (1983) Parallel Visual Computation, *Nature*, Vol. 306, pp. 21-26.

[13] Hummel, R. A., and Zucker, S. W., (1983) On the Foundations of Relaxation Labeling Processes, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 5, No. 3, pp. 267-287.

[14] Poggio, T., Torre, V., and Koch, C. (1985) Computational Vision and Regularization Theory, *Nature*, Vol. 317, pp. 314-319.

[15] Rissanen, J. (1983) Minimum-length description principle. *Encyclopedia of Statistical Sciences,*, Vol. 5, pp.523-527. New York: Wiley

[16] Leclerc, Y., (1989) Constructing Simple Stable Descriptions for Image Partitioning, *Intl. J. Computer Vision*, Vol 3, May 1989.

[17] Cheeseman, P., et. al (1988) Bayesian Classification, *Proc. AAAI-88*, pp. 607-611, St. Paul, MN

[18] Mohan, R., and Nevatia, R. (1989) Using Perceptual Organization to Extract 3-D Structures, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 11, No. 11, pp. 1121-1139.

[19] Terzopoulos, D., (1988) The Computation of Visible-Surface Representations, *IEEE Trans. Pattern Anal. Machine Intell.,* Vol 10, No. 4

[20] Besl, P., (1988) *Surfaces in Range Image Understanding*, Springer Verlag, New York

[21] Szeliski, R., (1989) *Bayesian Modeling of Uncertainty in Low-Level Vision*, Kluwer Academic Publishers, Boston

[22] Solina, F. and Bajcsy, R., (1990) Recovery of Parametric Models from Range Images: The Case for Superquads with Global Deformations, *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 12, Feb 1990

[23] Hopfield, J. J., and Tank, D. W., (1985) Neural computation of decisions in Optimization Problems, *Biological Cybernetics,* Vol. 52, pp. 141-152.

[24] Kim, D. Y., Meer, P., Mints, D., and Rosenfeld, A.; Robust Image Recovery by a Least Median Squares Technique, preprint, to be presented *Proc. Image Understanding and Machine Vision*; Optical Society; June 1989

[25] Press, W., et. al; *Numerical Recipes in C*; Cambridge University Press, New York; 1988

[26] Beaton, A. E, and Tukey, J. W.; The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics 16*; 1974

[27] Li, G.; Robust Regression. In D.C. Hoaglin, F. Mosteller and J.W. Tukey (Eds.) *Exploring Data, Tables, Trends and Shapes* John Wiley & Sons, N.Y.; 1985

[28] Koffka, K. (1935) *Principles of Gestalt Psychology* Harcourt, Brace & World, New York.

[29] Hochberg, J.E., and McAlister, E. (1953) A quantitative approach to figure 'goodness', *J. Experimental Psychology*, vol 46

[30] Leeuwenberg, E.L.J, (1971) A perceptual coding language for visual and auditory patterns *American Jour. of Psychology*, Vol 84

[31] Attneave, F. (1982) Pragnanz and soap-bubble systems: A theoretical exploration. in J. Beck (Ed.) *Organization and representation in perception* Erlbaum, Hillsdale, New Jersey.

[32] Wallace, C.S., Boulton, D.M., (1968) An Information Measure for Classification, *Comp. J.*, 11, pp. 185-195

[33] Fodor, J. (1983) *The Modularity of Mind*, MIT Press, Cambridge MA.

[34] Barrow, H.G., and Tenenbaum, J.M., (1978) Recovering Intrinsic Scene Characteristics from Images. in Hanson, A.R., and Riseman, E.M., (Eds.) *Computer Vision Systems* Academic Press, New York

[35] T. Poggio, E. Gamble, D. Geiger, and D. Weinshall (1989) Integration of vision modules and labeling of surface discontinuities *IEEE Trans. Systems, Man & Cybernetics*, December

[36] Adelson, E. (1990) Personal Communication

[37] Pentland, Alex, Part Segmentation for Object Recognition, *Neural Computation* Vol 1, 82-91