Sectored Snakes: Evaluating Learned-Energy Segmentations

Samuel D. Fenster and John R. Kender *
[fenster | kender]@cs.columbia.edu
Dept. of Computer Science, Columbia University, New York, NY 10027, USA

Abstract

We describe how to teach deformable models to recognize segmentation correctness based on user-specified criteria, and we present a method for evaluating which criteria work best. We present sectored snakes, a formulation that demonstrably improves upon regular snakes.

A traditional deformable model ("snake" in 2D), used for image segmentation, fails to find an object's boundary when the strongest nearby image edges are not the ones sought. But models can be trained to respond to other image features instead, by building their probability distribution. The implementor must then decide on which of many image qualities to teach the model. To this end, we show how to evaluate the efficacy of any resulting deformable model, given a sampling of ground truth, a model of the range of shapes tried during optimization, and a measure of shape closeness.

In the domain of abdominal CT images, we demonstrate such evaluation on a "sectoring" of a snake, in which intensity and perpendicular gradient are observed over equal-length segments. This specific set of qualities shows a measured improvement over an objective function that is uniform around the shape, and it follows naturally from examination of the latter's failures due to images variations around the organ boundary.

Keywords: *Segmentation and Grouping, Shape Representation and Recovery, Learning in Vision, Applications, Segmentation Evaluation*

*This work was supported in part by DOD/ONR MURI Grant N00014-95-1-0601, by the New York State Science and Technology Foundation, and by ARPA Contract DACA-76-92-C-007.

Contents

1 Introduction

In [Kass *et al.*, 1987], the authors introduced a flexible shape model that moved until it maximized some combination ("energy") of its smoothness and the cumulative strength and proximity of nearby image edges. But there was no "right" way to set the parameters of this combination [Boult *et al.*, 1994].[1] Furthermore, it was not established how effective an objective function based on those particular quantities was in the first place. Since then, specific other features have been tried in various domains, *e.g.* matching intensity profiles at feature points [Cootes *et al.*, 1993], matching multiscale intensity and gradient at feature points [Baldwin, 1997], and centering within a region of classified pixels [Davatzikos & Prince, 1995]. But few, if any, comparative studies within a domain have been done.

Let us consider what quantity should be optimized when seeking a shape in an image. To find the boundary of a specific kind of object in a cluttered scene, a shape model should know what it looks like, in a way that differentiates it from nearby object boundaries. If this is not known *a priori*, analytically or otherwise, then the model should *learn* what it looks like. That is, through training it should learn the likeliest values of selected image features associated with the boundary of that kind of object. Which features are selected determines, through training, what function will be optimized in order to find an object boundary. Some feature choices will work better than others.

In this paper, we introduce a general method to evaluate how well a given kind of training will

[1][Lai & Chin, 1995] used a minimax criterion to balance shape and image components of their objective function in a manner that had nothing to do with what object was sought in the domain.

perform in a given domain. We then introduce the sectored snake, a particular improvement on the most straightforward kind of training, which is itself an improvement on standard snakes.

1.1 Deformable models and the need for improvement

A *deformable model* is a description of a shape which is iteratively adjusted until it best matches what is depicted in an image. "Best" is measured by an arbitrary real-valued objective function, or "energy." The process of minimizing this energy is sometimes called a "force." In the case of a contour in a 2D image, the model is called a *snake* [Kass *et al.*, 1987]. As a method of segmentation it has many advantages. It can fit a continuous shape to discrete image data; it accommodates shape models that can change locally to conform to image details and, conversely, can constrain global shape; prior shape expectation (local and global) can be incorporated in the objective function. Thus, it can be robust to locally bad data, yet can find local shape where data exists. It can respond to general image-processing cues (edges) or domain-specific information such as particular colors or textures.

There are domains in which traditional snakes, attracted to strongest or closest image edges, fail. The strongest edges are often not the edges of the object we seek. This is true in the domain in which we are carrying out our studies—CT images of lower abdominal organs, which are pressed up against similar organs and brighter, stronger-edged bone (Fig. 1). This is not just a problem of finding suboptimal local minima—in such cases, the wrong object satisfies the objective function better than the right one does. Researchers have formulated alternative image forces to get around the problem, without formally testing their properties in the domain, and without having some

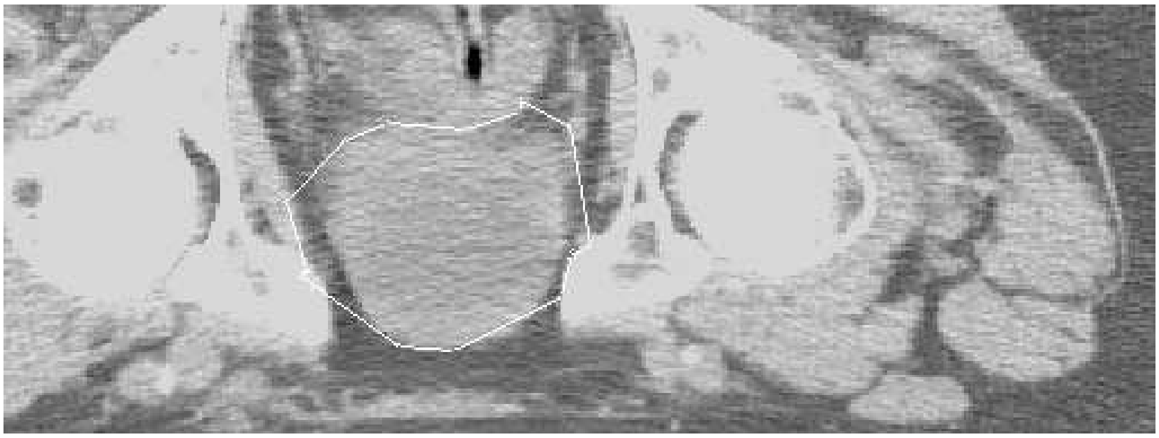

Figure 1: *An organ (bladder) next to stronger-edged bone (pelvis). A traditional snake, attracted to strongest edges, incorrectly clings to bone. The solution is to teach it to recognize the qualities of the correct boundary.*

basis to compare alternative forces.

1.2 Steps toward a solution

If the structure whose contours are sought does not have features (*e.g.* brightness, edge strength) that are more extreme than those of all neighboring objects, the energy being minimized will need its parameters tailored to some domain-dependent intermediate value. That is, it cannot merely achieve an optimal value when the contour lies on, say, the strongest edges or the brightest region—it must know the *correct* feature strength to reward. This must be found through statistical measurement, which is to say, training. This requires a ground truth training set, which is a collection of images, each with the desired contour correctly specified. Such training on image features has been done in [Cootes *et al.*, 1993], whose model learned a knD Gaussian of intensities along line segments of n pixels perpendicular to a shape boundary at k feature points. Within a similar framework, [Baldwin, 1997] used a Gaussian distribution of multiscale intensity and gradient at a few points around many organs simultaneously in a human cross-section. Finally, [Boes *et al.*, 1995]

Figure 2: *Image variability. These are bladders of different patients. There are parts of their boundary that vary greatly between images, such as the top, pressed against the highly variable rectum. Other parts of the boundary may have very predictable appearance, like the sides, which are always near pelvic bone, and the seminal vesicles, on either side of the top.*

uses a 3D model incorporating the observed likelihood of nearby edges belonging to the liver or not.

The qualities that training teaches the model to seek may not be uniform everywhere on its boundary. For instance, in abdominal CT scans, the bladder may be in contact with the pelvis on the sides, seminal vesicles (in men) at two regions near the back, the (highly variable) rectum directly behind, and tissue fluid elsewhere. Thus, its boundary characteristics vary with position (Fig. 2). This calls for a spatially varying objective function. Furthermore, they may have differing degrees of variability at different places, so the function's sensitivity should vary, providing robustness where wide variation is expected. We will present sectored snakes, which address these needs within the training framework.

1.3 Measuring effectiveness

With such a choice of objective functions, we must be able to compare and evaluate their performance, which is domain-specific. Contrary to intuition, the distribution of an objective function's values for *correct* contours gives no information about its goodness for segmentation. Consider that, whatever this distribution, if incorrect contours have the *same* distribution of values, the function cannot guide a contour to the correct shape. Qualities consistent across the body of ground truth are useless if they are also unchanged for wrong contours. We must also look at nearby *incorrect* shapes to assure that an objective function responds to increased correctness by getting closer to optimal. This criterion depends on the image environment and variability near each contour. We examine an energy function's behavior for incorrect shapes, by generating such shapes from ground-truth shapes in a sample of domain images.

1.4 The rest of this paper

The next section describes the domain in which we performed our experiments. We then describe a general framework for training deformable contour models. Next, we describe how to evaluate a deformable model, trained or otherwise. We then describe sectored snakes. Finally, we present experimental results and evaluations.

2 Domain: Radiation Treatment Planning

In our domain, object contours in the human abdomen are needed to plan radiation treatments which will destroy tumors in the prostate. In such treatments, shaped beams of radiation converge on the tumor so as to give it maximum exposure while giving surrounding healthy tissue minimal exposure. Before beam shape and direction can be optimized, the layout of all nearby organs must be known. Each of these is outlined manually, slice by slice, on CT scans, sometimes with the help of simple edge-following algorithms. This is a slow and costly process. A repetitive task on which human life depends could use some verifiably accurate machine assistance. Automated segmentation would help greatly, but, because the image is crowded and complex, it is difficult.

2.1 "Ground" truth

Thus, there is a large body of hand-segmented ground truth in this domain. This data may be used to train deformable contour models and to test the accuracy of the resulting models. For our experiments, we concentrated on median slices of the bladder. Although this was a small subset of available images, it maintained the critical temporal and interpatient variations we were interested in. The possible proximity of a tumor gave our training data the same incidence of organ abnormalities as would be seen in the application of training results in the domain. Our images were from a study in which patients had four images taken at different stages of treatment.

Upon investigation, it appeared that the "ground truth" available to us was not drawn very accurately (Fig. 3); it is often about six pixels away from the visible organ boundary, in a 512×512 image. One might think that this would make it easier for our segmentation algorithms to match human performance, which currently suffices in this application. But a poor objective function in a snake segmentation does not necessarily yield small inaccuracies within the range of ground truth variation; rather, it will produce contours that follow entirely the wrong organ in places. And inaccurate data is a disadvantage for us because it cannot provide training that will result in accurate segmentations.

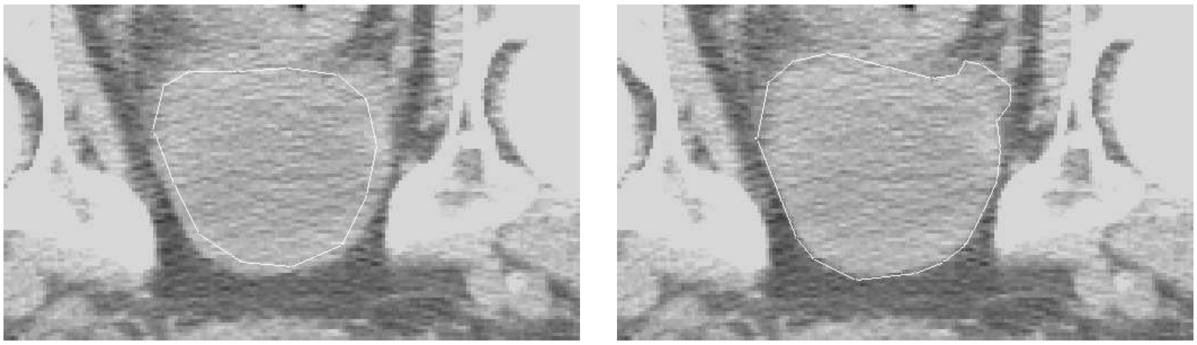

Figure 3: *Medical personnel did not draw contours to pixel accuracy (left). Learned segmentation can be no better than the training data, so we redrew the contours we trained on (right).*

2.2 Better ground truth

To get accurate training data, we drew our own contours, with agreement from doctors that we could do better than the ones they routinely drew. (Economics and time constraints prevent them from doing better.) For greatest consistency, we trained and tested contours from the organ's middle slice (or two, if there were an even number), where each organ had four to 14 slices (averaging between nine and ten). Up to four 3D images from each of seven patients yielded 24 organ images in 3D, upon whose middle slices we drew 36 high-quality bladder contours.

3 Learning

3.1 Why learn?

Accurate contours of known objects in cluttered images are needed in many applications, including medicine and manufacturing. No segmentation method that does not use domain knowledge will be able to distinguish the sought object from neighboring ones. Since many real-world objects, including human organs, do not have qualities that are known *a priori* or analytically, these qualities must be measured. If image qualities are to guide segmentation, it is the image features of these objects which should be measured. These qualities will generally have distributions, not precisely repeated values. These distributions must be recovered from observations, *i.e.* training images with known contours. Once they are known, the desired object may be found in a new image by finding the shape whose relationship to the image is likeliest, according to the known probability distributions.

Traditional snakes, by contrast, are based on aligning a contour with image edges of maximum strength. But machine vision cannot always be based on generic image features. Without knowledge of the domain, an implementor cannot always know what the right answer even is. This ambiguity is what foils traditional snakes in some domains. Thus, a snake must be taught about correct and incorrect shapes. Although there has been plenty of work on learning the prior probability of a shape, *e.g.* [Kervrann & Heitz, 1994, Lai & Chin, 1995], there has been little on learning a model of shape likelihood given an image.

3.2 Formalizing how to learn

Here, we formalize the notion of learning a general deformable model objective function based on arbitrary shape and image features. Before learning the probability distribution of image qualities associated with a contour, we must select two things: which image qualities to observe, and what model of a probability distribution to fit to them.

To approach the problem, we formalize it in the following way (Fig. 4). We have an image **I** in

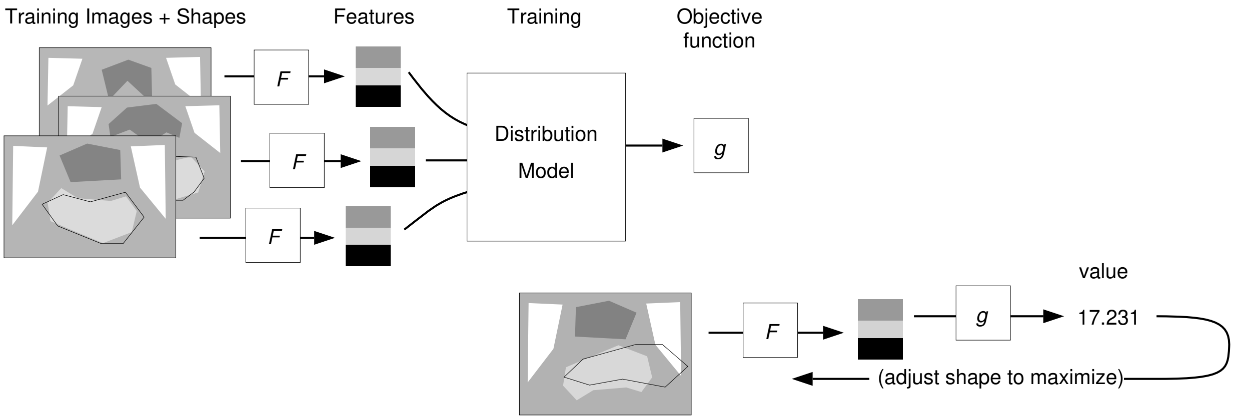

Figure 4: *To train a deformable model, we need to decide on a set of features, extracted from an image and shape by function F. We then need a continuous model of the density of feature values which might be observed. Features are extracted from correct image/shape pairs and their distribution is modeled as objective function g. Now, to find the desired shape in a new image, we find a contour in that image which produces the feature values that g says are likeliest.*

which a shape $\mathbf{S}$ may or may not be a good approximation to the boundary of the object we seek. We have decided to base our guess as to the correctness of $\mathbf{S}$ on quantities which say something about it and its relation to $\mathbf{I}$. The function $F(\mathbf{I}, \mathbf{S})$ extracts this tractable set of quantities from its large input.

The values of these chosen features will, hopefully, indicate whether a contour is correctly placed on the boundary of the desired object in an image. To know what values of F are "good," we must learn the distribution of these values for correct shapes. A probability density function (PDF) can represent this. It will be a real-valued, continuous function $g(F(\mathbf{I}, \mathbf{S}))$, recovered from a finite set of training data, representing the density observed in that training set in the region of any given vector of feature values $F(\mathbf{I}, \mathbf{S})$. It will approximate the probability $\mathrm{P}(F(\mathbf{I}, \mathbf{S}))$. Before we can recover such a distribution g, we must have some finitely parameterizable representation for it.

Once we have picked F and a model to represent g, we can train and use a deformable model.

First we must get ground truth: a set of images $\{\mathbf{I}_1, \ldots, \mathbf{I}_n\}$ of the desired object, with its correct outlines in each image, $\{\mathbf{S}_1, \ldots, \mathbf{S}_n\}$. Features from each pair, $F(\mathbf{I}_i, \mathbf{S}_i)$, are the data from which the parameters of a continuous distribution g are recovered. The likeliest parameters are those defining a distribution that maximizes the joint probability of observing the training set's features, $\mathrm{P}(F(\mathbf{I}_1, \mathbf{S}_1) \wedge \ldots \wedge F(\mathbf{I}_n, \mathbf{S}_n))$. Since the training data were presumably drawn independently, we must thus find the function g that maximizes

$$\prod_{i=1}^{n} g(F(\mathbf{I}_i, \mathbf{S}_i))$$

Once training has given us g, a deformable model can segment an image $\mathbf{I}$ by finding the optimal shape $\mathbf{S}^*$. (The optimization method used is not addressed here):

$$\mathbf{S}^* = \arg\max_{\mathbf{S}} g(F(\mathbf{I}, \mathbf{S}))$$

$g(F(\mathbf{I}, \mathbf{S}))$ approximates $\mathrm{P}(F(\mathbf{I}, \mathbf{S}))$. For well-chosen F and model for g, the shape maximizing this probability will also maximize $\mathrm{P}(\mathbf{I} \wedge \mathbf{S})$, and thus also $\mathrm{P}(\mathbf{S} \mid \mathbf{I})$, since $\mathrm{P}(\mathbf{I})$ is constant as $\mathbf{S}$ varies. In a Bayesian approach, this is the maximum-likelihood shape.

A well-chosen feature function F is one for which correct shapes will have different feature vectors than incorrect ones. (Fig. 5 illustrates what might make a feature function do poorly.) An adequately chosen probability model is one for which the recovered g is flexible enough to have high values only for feature vectors commonly produced by correct shapes.

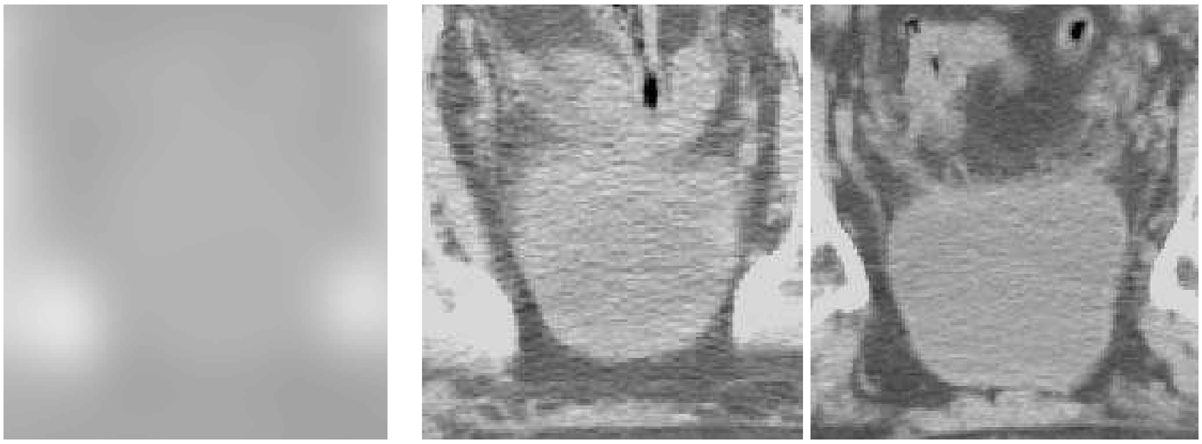

Figure 5: *For any given domain, there are image qualities that, when learned, will not successfully guide a contour to an object boundary,* i.e. *training will not produce a good objective (energy) function.* Left: *Boundary intensity or gradient strength at too coarse a scale will not distinguish a correct contour from a wrong one because distinguishing image detail is gone.* Middle and right: *If boundary intensity varies more between images than between the object and its neighbors, no simple intensity-based objective function can respond better to the right boundary than to the wrong one.*

3.3 A basic implementation of trained snakes

In our initial trained segmentation of bladder boundaries in CT scans, we made the following choices. $\mathbf{S}$ was a closed polyline $\mathbf{S}(u)$. Image quantities were observed at a scale s, using a Gaussian-blurred image $\mathbf{I}_s = \mathbf{I} * G^2_{0,s}$. The shape's relation to the image was summarized by image intensities, $\mathbf{I}_s(\mathbf{S}(u))$; and by directional image gradients normal to the shape, $\mathbf{S}^{\perp}(u) \cdot \nabla \mathbf{I}_s(\mathbf{S}(u))$, where $\mathbf{S}^{\perp}(u)$ is the unit normal to $\mathbf{S}(u)$. These features, measured along $\mathbf{S}$ at one-pixel intervals, can be considered the output of $F(\mathbf{I}, \mathbf{S})$.

We chose perhaps the simplest model of the distribution of our features—a multidimensional Gaussian. If the incidence of our chosen feature values in our domain is unimodal, a Gaussian may model it well enough. Specifically, we assumed independent, identically-distributed (I.I.D.) values of intensity $\mathbf{I}_s(\mathbf{S}(u))$ and directional gradient $\mathbf{S}^{\perp}(u) \cdot \nabla \mathbf{I}_s(\mathbf{S}(u))$ at all points along $\mathbf{S}$. Thus, training recovered the parameters of two Gaussian distributions, $N(\mu_{\mathbf{I}}, \sigma_{\mathbf{I}})$ and $N(\mu_{\nabla}, \sigma_{\nabla})$. The joint probability of both these quantities at every point around the contour was the modeled probability of observing those features on a shape $\mathbf{S}$ in an image $\mathbf{I}$, *i.e.* $\mathrm{P}(F(\mathbf{I},\mathbf{S}))$. This is a product of Gaussians; its negative log is

$$E_{\mathbf{I}} = \oint \frac{\left(\mathbf{I}_s(\mathbf{S}(u)) - \mu_{\mathbf{I}}\right)^2}{\sigma_{\mathbf{I}}^2} du + \oint \frac{\left(\mathbf{S}^{\perp}(u) \cdot \nabla \mathbf{I}_s(\mathbf{S}(u)) - \mu_{\nabla}\right)^2}{\sigma_{\nabla}^2} du$$

$E_{\mathbf{I}}$ is the "image energy." Minimizing this "energy" is equivalent to maximizing $\mathrm{P}(F(\mathbf{I},\mathbf{S}))$. We use straightforward gradient descent minimization over shapes $\mathbf{S}$, so we must take the derivative of $E_{\mathbf{I}}$ with respect to the parameters of $\mathbf{S}$.

4 Sectored Snakes: A new synthesis

We will see that in a cluttered domain, even the simple training described above provides a much better-behaved objective function than does the traditional snake energy. The latter rewards contours for sitting on the strongest image edges. A trained contour, on the other hand, seeks the relationship to image data that was observed in the training set. Training will only work if some aspect of this contour-image relationship is consistent across the training set (and is different for incorrect contours).

In using snakes trained to seek a particular boundary brightness and edge sharpness, we noticed that, unsurprisingly, a snake would behave badly where boundary qualities were locally different. Since these spatial boundary variations are often consistent across images of a particular domain object (Fig. 2), it made sense to formulate a snake that learned what to seek separately on separate portions of its length. In this, we followed [Cootes *et al.*, 1993], although they trained a small number of preselected feature points. We seek an accurate boundary everywhere, so we wish to respond to image data everywhere along the contour, rather than just at a few points.

We thus divided the contour into a fixed number of equal-length regions (sectors), each with separate training (Fig. 6). Not only would this allow the snake to be attracted to different conditions where different conditions were expected; it also would allow stronger objective function response in regions where there is less variability in conditions between images.

To get regions that roughly correspond among training contours, and between training data and images to be segmented, an origin for the start of the first sector, going clockwise, had to be selected. In many domains, including ours, images and the objects therein may be expected to have a consistent orientation. Thus, we chose "twelve o'clock" to be the highest contour point above the center of mass of the contour (which is quicker and easier to calculate than the center of mass of the enclosed area).

When we minimized our simple, unsectored image energy, we were finding a maximum-probability shape assuming a joint I.I.D. Gaussian distribution of image intensities and perpendicular gradients (at some scale) observed around the contour. In a snake with M sectors, this energy is modified to reflect one of M Gaussian distributions, based on position along the contour. Thus, we have a function $k(u)$ which maps the shape parameter u to a sector, $1, \ldots, M$. Instead of a single pair of Gaussians, $N(\mu_{\mathbf{I}},\ \sigma_{\mathbf{I}})$ and $N(\mu_{\nabla},\ \sigma_{\nabla})$, we have a pair for each sector: $N({}_k\mu_{\mathbf{I}},\ {}_k\sigma_{\mathbf{I}})$ and $N({}_k\mu_{\nabla},\ {}_k\sigma_{\nabla})$.

Our image energy still has a scale parameter s so that the image qualities it responds to are in

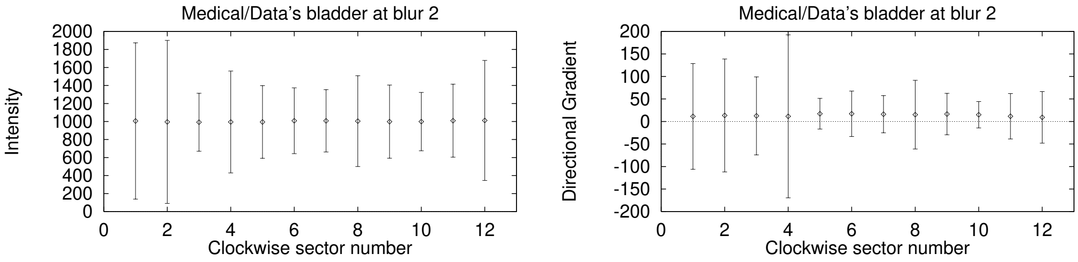

Figure 6: *Each sector of a sectored snake has separately learned a joint Gaussian of image intensity and gradient in the normal to the snake direction. The intensity bars (left plot) are tallest in regions 12, 1 and 2, corresponding to greatest intensity variability at the top of the bladder. A snake with this data would not be as strongly attracted to any particular intensity value at the top as elsewhere, since it would "know" that a wide range of intensities occurred in that region, in the training data.*

a Gaussian-blurred image $\mathbf{I}_s = \mathbf{I} * G^2_{0,s}$. So our sectored image energy $E_{\mathbf{I}}$ of a contour $\mathbf{S}(u)$ in an image $\mathbf{I}$ at scale s is:

$$E_{\mathbf{I}} = \oint \frac{\left(\mathbf{I}_s(\mathbf{S}(u)) -_{k(u)} \mu_{\mathbf{I}}\right)^2}{_{k(u)}\sigma^2_{\mathbf{I}}} du + \oint \frac{\left(\mathbf{S}^{\perp}(u) \cdot \nabla \mathbf{I}_s(\mathbf{S}(u)) -_{k(u)} \mu_{\nabla}\right)^2}{_{k(u)}\sigma^2_{\nabla}} du$$

where $\mathbf{S}^{\perp}(u)$ is the unit normal to $\mathbf{S}(u)$. The probability we are modeling, $\mathrm{P}(F(\mathbf{I}, \mathbf{S}))$, is proportional to $e^{-\ln E_{\mathbf{I}}}$.

The number of sectors was picked large enough to capture the number of different types of regions around the object we sought. In our tests, this was the bladder, which has pelvic bone on each side and seminal vesicles (in men) and rectum in back. If the training data were infinite, more sectors would always be better, because their training would approximate a continuously varying probability distribution of image qualities around the shape boundary, and would capture arbitrarily small variations. But since the training data is finite, more sectors also means that each sector gets less exposure to a statistically representative variety of surroundings in the image, and thus poorer training. We chose twelve sectors based on the sizes of structures that seemed consistently positioned around bladder boundaries. Further work may examine the effect of varying the number; sectors of unequal length; and continuous parameterizations of the spatially varying probability distribution. For our initial tests, we wanted to see if our intuition was borne out that even a simple sectoring of a snake in a consistent domain would improve our segmentation over that of spatially uniform training.

The next section addresses how to rate the effectiveness of this (or any other) formulation.

5 Presenting: How to Test Goodness of Learning

The implementor of a deformable model can choose from a wide variety of objective functions. If the function is to be derived from training data, this variety is equivalent to the selection of features to model the distribution of, and of distribution models to use for each. It is essential that we be able to judge the efficacy of any given function. If the function results from training, we are judging the feature set and distribution model.

Thus far, no methodology has existed for this. The standard procedure is to run the entire deformable model segmentation procedure on a series of images, comparing each resulting shape to the ground truth for that image. But the results of this method only characterize the cumulative effect of: selection of the contour's initial position; the objective function; and the optimization method to find the best nearby contour. It cannot test the effectiveness of their separate contributions to segmentation. Generally, each image is tried with only one initial contour. This method is inadequate to evaluate objective functions, whether derived from training or not.

If we want to find a suitable objective function, or suitable image qualities to train on, we must evaluate the function over a range of initial positions and of intermediate configurations that might be tried during optimization. For optimization to succeed, one condition must hold over this range of shapes: *The function must be optimal for the correct shape.* If this condition is not satisfied, any optimization method will pick the wrong shape. If gradient-descent optimization is used, a more stringent condition must be met: *The function must become more optimal as a shape gets closer to correct.* If this condition is satisfied, gradient descent will approach the correct shape. If it is not satisfied, the optimization may pick an incorrect local minimum.

To statistically characterize a function's behavior, we randomly perturb the ground-truth shape within an image to produce more- and less-incorrect contours. A perturbation model cannot densely cover the (very large) space of possible nearby shapes; but it can be used to reject objective functions in which the subset of shapes it generates *does* reveal flaws. Such a test is much more complete than running just one segmentation on each image; much more efficient than running a large number of complete segmentations; and, as a test of a single (key) component, more informative than only seeing the result of the entire optimization process for segmenting an image. In the next section, we describe the model of random perturbation we used: normally distributed first-order deformations of the contour.

Given an image, we must compare the objective function's valuings of perturbed shapes (in the absence of ground truth) to some measure of each shape's distance to the known correct boundary (ground truth). We are checking that known worse shapes have worse objective function values. Thus, we need a measure of how far a shape is from another (ground truth) shape. We choose such a measure, chamfer distance, following the next section.

We must evaluate the function over a range of images, and, for each, over a range of contours that encompasses all that might be tried in the search for the correct contour. If a number of different objective functions are being evaluated, perturbed contours and their distances from ground truth need only be generated once. We tested six different objective functions (three different models, each at two scales) using 1000 parameterized perturbations of known bladder outlines in each of 36 different 2D images.

The relationship of objective function to known shape correctness (distance to ground truth) over the expected range of shapes tells us whether the function will perform well. For each test image, this relationship is described exactly by a scatter plot of these two quantities for many perturbed ground truth contours. We wish to test whether any deformation of a shape that decreases its energy actually brings it closer to ground truth. We get the strongest assurance of this from scatters which are most tightly clustered around some increasing function. There are various ways this relationship may be quantified. One is by calculating the correlation coefficient—the closer it is to +1, the better. Another key statistic is the number of contours having energies below that of the ground truth, thus falsely indicating that they are better solutions than the correct one. Ideally there would be none of these. Whatever statistic is chosen, one must combine its values for many images. We now describe the implementation decisions we made to realize these results.

5.1 Range of contours: A statistical approach

We get a statistical sampling of shapes around a ground-truth shape by perturbing that ground-truth shape. Since we cannot densely cover the high-dimensional neighborhood of all nearby or similar shapes, we use a set of parameterized perturbations which will expose expected kinds of flaws in the objective function(s) we are testing.

The perturbations we used for objective function evaluation were a combination of translation and of scaling independently around two randomly chosen orthogonal axes centered on the contour's centroid.. The translations were normally distributed with a standard deviation of 5 pixels; the logs of the scale factors were normally distributed with a standard deviation of $log(1.1)$, or 10% stretch/shrinkage.

5.2 A measure of distance to known ground truth

For each perturbed contour, we need to record its energy (objective function value) and its distance from the correct shape. Possible choices for the latter included the Hausdorff distance between the shapes, the chamfer distance, the area of the symmetric difference of the enclosed regions, and others.

The Hausdorff distance, since it it depends on minima and maxima of distances between points on two contours, is not robust to small protrusions in either shape. The symmetric difference of areas was inefficient to calculate. We chose the chamfer distance between two boundaries [Borgefors, 1988], which is the average over one shape of distance to the closest point on the other. To calculate this, one does a distance transform on an image containing one shape, then averages the distance values intersected by the other. This measure is asymmetric, but (in our domain, at least) has a high correlation (0.977) between $d(a, b)$ and $d(b, a)$. (This correlation is the mean for our test set of 36×1000 perturbed contours; its standard deviation is 0.005.) If a commutative measure were needed, one could use the average of the two directions. (The triangle inequality of a formal metric is not needed here.) The asymmetric chamfer distance gives a result of zero in the pathological case where two curves coincide except for a protuberance on the one to which the distance transform is applied. It is, however, more robust to this pathology than the Hausdorff distance, and the forward-reverse correlation over our test set shows that it did not arise much. We have recently become aware of a possibly better measure of distance between curves, [Chalana & Kim, 1996].

5.3 Putting it all together

Once we have a model of how to perturb ground-truth contours, and we have a measure of the distance between two contours, we can proceed to characterize the goodness of a deformable model's objective function. This goodness can be summarized in a plot of shape incorrectness *vs.* function value (Fig. 7).

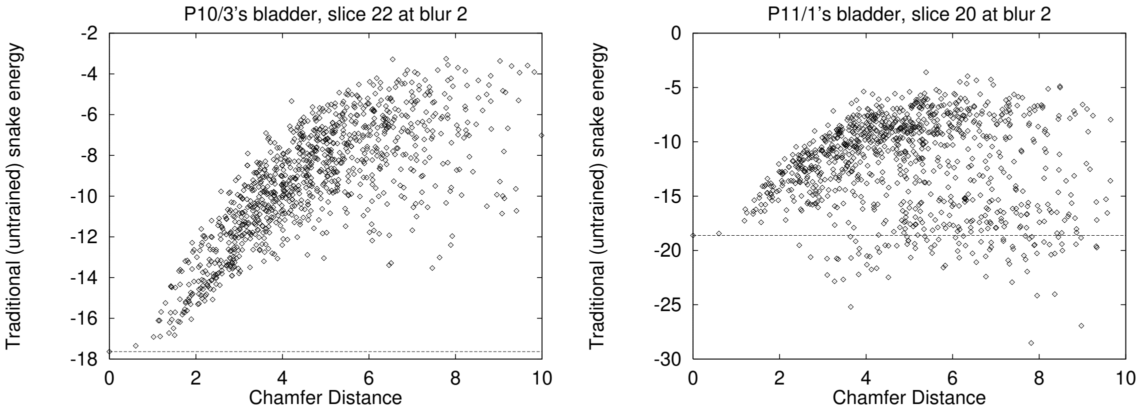

Figure 7: *Examples of plots showing how effective a deformable model's objective function will be in a domain. Given an image and the correct contour of a shape in it, a large set of perturbations of the shape are generated. Each perturbed contour is a point in a scatter plot of difference from correct shape* vs. *energy. A good objective function will display strong correlation with shape correctness, and will have no "false positives," contours with lower energy than the ground truth (dotted line).* Left: *a high-correlation (0.73) plot with no false positives.* Right: *A low-correlation (-0.04) plot with 8.8% false positives. Each plot characterizes an energy function's behavior in a single image. To rate the function's performance in a domain, statistics of plots from many images must be aggregated.*

6 Results

To the best of our knowledge, this is the first formal measurement and comparison of the effectiveness of deformable model energy functions. As described in section 2.2, our training data were 36 hand-outlined 2D images of bladders from 24 abdominal 3D CT scans of seven patients. More data are being made available to us.

Training and segmentation was done for three different snake formulations: the sectored snake described in section 4; the simple trained model described in section 3.3; and a traditional, untrained snake whose energy is gradient strength traversed by the contour, at some scale: $E_{\mathbf{I}} = -\oint ||\nabla \mathbf{I}_s(\mathbf{S}(u))||du$. Each of these was tested at two scales, $s = 2$ and $s = 4$, for a total of six different objective functions.

We evaluated each of these energy functions, using the procedure described in section 5, with

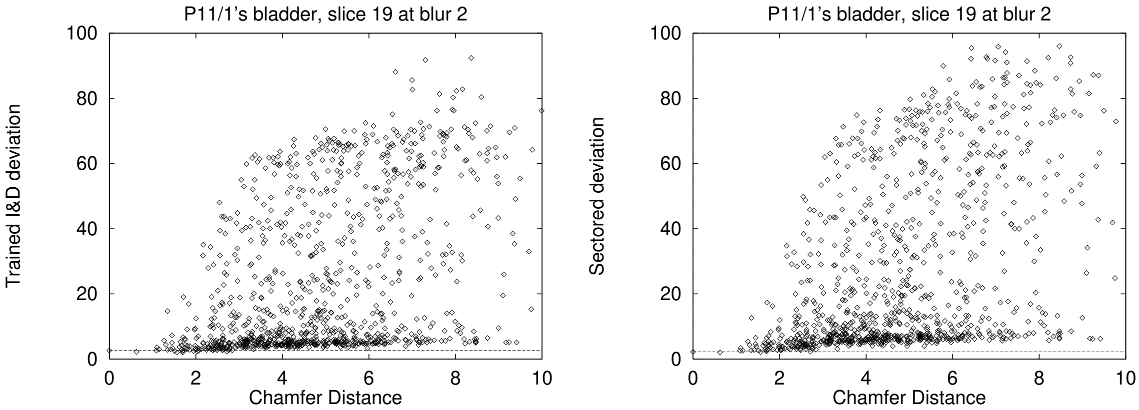

Figure 8: *Two objective functions—the result of unsectored and sectored training on the same dataset—applied to contours in the same image. The former has a correlation of 0.43 and 11 false positives; the latter has a correlation of 0.48 and 3 false positives (below the dotted line).*

1000 perturbations of the ground truth contour in each of the 36 slices. Below we look at two statistics of the results. The first is the degree of correlation between a contour's image energy and distance from the image's ground truth contour. The second is the number perturbed contours whose image energy is lower than that of the unperturbed ground truth contour ("false positives").

6.1 Correlation

Our aim is to check whether contours that are further from ground truth have consistently higher (less optimal) energies. One guarantee of this is if the scatter plot of distance d from ground truth *vs.* energy E is tightly clustered around some increasing function. A high linear correlation coefficient $\sum(d_i - \bar{d})(E_i - \bar{E})/n\sigma_d\sigma_E$ is an indicator of this, although a low value does not preclude it. This statistic is shown, for each energy function, in Table 1.

Snakes with sectored training had slightly, but not significantly, higher correlations than unsectored trained snakes, at both image scales. Coarser (blurrier) image scale increased correlation by a small amount. But most importantly, both of the trained formulations demonstrated a much higher correlation than traditional snakes.

Objective function	Scale	Average Corr	Deviation of Corr
Sectored and trained	2	0.56	0.12
	4	0.61	0.13
Unsectored but trained	2	0.55	0.14
	4	0.58	0.14
Untrained (traditional)	2	0.15	0.33
	4	0.21	0.37

Table 1: Degree of correlation between contour image energies and chamfer distance to correct shape. In our domain, the sectored snake's energy is slightly more correlated to shape correctness than the unsectored snake, and is much better than a traditional snake. Though correlation is higher at a blurrier scale, the following tables show that the bigger blur has problems.

6.2 "False positives"

An energy function is guaranteed to produce some incorrect segmentations if it is smaller for some close but incorrect contours than it is for the ground truth. Such "false positives" occurred far less often with sectored snakes than unsectored snakes (*e.g.*, Fig. 8). Unsectored trained snakes, in turn, performed much better than traditional snakes. The average incidences of false positives per image for each energy function are in Table 2.

Although the coarser-scale energy functions showed correlations with distance to ground truth that were just as good as at the finer scale, they have many more false positives—a factor of ten, for the two trained formulations. This disqualifies scale $s = 4$ for use in segmentation in this domain. The correlation measure is equally good at the two scales, while the false-positive measure is very bad for the coarse scale.

Objective function	Scale	Average FPs	Deviation
Sectored and trained	2	0.6%	0.8%
	4	5.8%	6.4%
Unsectored but trained	2	1.1%	1.6%
	4	10.9%	10.9%
Untrained (traditional)	2	15.4%	14.5%
	4	23.2%	16.9%

Table 2: Incidence of "false positives" (contours with objective function values better than correct shape) for different objective functions. The sectored snake decisively outperforms the snake with unsectored training in this image domain. More blurring causes each of the three function types to make many more mistakes.

Table 3 shows what portion of the 36 image slices met progressively more relaxed standards (number of false positives allowed), for each energy function. For instance, 72% of the image slices had less than 2 in 1000 bladder contour perturbations with lower sectored-snake energies than the truth. Only 28% of the images had that few glitches with the traditional snake energy.

Objective function	Scale	no FPs	FPs $<$ .2%	FPs $<$ 2%	FPs $<$ 5%
Sectored and trained	2	53%	72%	94%	97%
	4	8%	11%	56%	72%
Unsectored but trained	2	42%	72%	89%	92%
	4	8%	14%	31%	56%
Untrained (traditional)	2	14%	28%	33%	42%
	4	14%	22%	31%	33%

Table 3: Percentage of 36 images in which fewer than the specified percentage of 1000 perturbations were false positives.

6.3 Discussion

The evaluations show us that in our domain of cluttered images, the enhancements we made to snakes were necessary to get good segmentation results. The simple statistics we extracted from energy-distance plots showed that in this domain, a standard, untrained snake would have failed most of the time. The data also provide evidence that heterogeneous training (sectoring) reduces the learned objective function's false positive rate significantly over simple unsectored training.

These tables show much better performance for energies based on the fine image scale than for those on the coarse scale. Unfortunately, when using gradient-descent optimization, coarser scales are necessary to get a nonzero gradient (derivative of energy with respect to shape parameters). An unblurred energy quickly reaches zero (plus noise) as a contour departs from the correct shape, so gradient descent will move an incorrect contour toward the solution. Since blurring seems to often prevent energy functions from being able to distinguish right and wrong boundaries, and yet blurring may be necessary for gradient descent optimization, gradual deblurring is indicated—a standard robust optimization technique recommended in [Kass *et al.*, 1987].

One should not reject coarse scale out of hand, though. The reader should remember that these results, and all ratings of segmentation algorithms, are domain-specific. For a domain in which object boundaries were less accurately drawn, or for other reasons not drawn on visible edges, a coarser scale might in fact provide a better correlation between a contour's correctness and image properties along it. This is because image properties on the boundary would be "visible" further away if they were blurred.

7 Conclusion and Future Plans

We have presented a systematized framework for the training of deformable models based on any chosen shape and image features and objective function model. We hope that this framework will allow unified analysis of the diverse deformable model formulations that have existed up to now. Within this framework, we have presented two results in the segmentation of imagery by means of deformable models.

The first is sectoring, a modification of snakes that incorporates the virtues of local adaptability with the known advantages of prior work. This synthesis appears to be especially useful in imagery characterized by "clutter." The method appears to be quantifiably better, at only a small increment in complexity. Its most telling advantage is that standard approaches would be provably inaccurate in these domains.

The second is a methodology that allows us (and others) to quantify how good deformable models are, including sectored snakes. We have tested this approach on what we hope is a believably large dataset of images, using novel but natural criteria for "goodness," and a means of anticipating segmentation failures that improves current practice in both efficiency and thoroughness.

Taken together, we believe these constitute a more robust and more quantitative approach to a difficult but critical image task. This approach can quite easily be shown to improve not only on human results, but on existing computerized approaches.

Since the method is really a class of methods, perhaps the major contribution is that it permits knowledgeable experimentation over many image features, allowing researchers to compare results more accurately, determine causes of objective function failure, and devise more precise measures of model accuracy and reliability.

We intend to extend the work in several mutually supportive directions:

- We will attempt at least one other domain of imagery, selected from the manufacturing domain, to see how well the experience we have gained applies, and what, if any, domain-dependent assumptions we have unwittingly designed into our measure of goodness.

- The class of image features is easy to extend beyond intensity and gradient measures; in addition to higher-order measures such as texture, this method easily accommodates the use of non-standard measures such as boundary placement relative to distant landmarks.

- The definition of a "sector" needs to be explored, particularly with respect to whether image statistics can specify some "natural" sectoring more robustly than our 12 points of the clock face. Also to be explored are continuously spatially varying training, as opposed to discrete sectors; and the existence of dependence between expected features at different locations.

- We must check if some other probability models fit the distributions of some features better than a Gaussian. For instance, a multidimensional spline could be fit to training feature density data. This, of course, requires even more experimentation on more hand-drawn ground truth.

- In the most ambitious setting, this method can allow even dynamic and automatic selection of image features, sectoring, and feature distribution model selection, as a kind of "meta-learning."

References

[Baldwin, 1997] Baldwin, B. (1997). *Multiscale Snakes*. PhD thesis, Courant Institute of New York University. To appear.

[Boes *et al.*, 1995] Boes, J. L., Meyer, C. R., and Weymouth, T. E. (1995). Liver definition in CT using a population-based shape model. In: *Proc. 1st Int. Conf. on Computer Vision, Virtual Reality, and Robotics in Medicine (CVRMed '95)* pp. 506–512, Nice, France: Springer-Verlag.

[Borgefors, 1988] Borgefors, G. (1988). Hierarchical chamfer matching: A parametric edge matching algorithm. *IEEE Tran. on Pattern Analysis and Machine Intelligence,* **10** (6), 849–865.

[Boult *et al.*, 1994] Boult, T. E., Fenster, S. D., and O'Donnell, T. (1994). Reinterpreting physically-motivated modeling. In: *Proc. ARPA Image Understanding Workshop* , Monterey, CA, USA: Morgan Kaufmann.

[Chalana & Kim, 1996] Chalana, V. and Kim, Y. (1996). A methodology for evaluation of image segmentation algorithms on medical images. In: *Image processing* volume 2710 of *Medical imaging* pp. 178–189, SPIE–the Int'l Society for Optical Engineering Newport Beach, CA, USA: SPIE.

[Cootes *et al.*, 1993] Cootes, T., Hill, A., Taylor, C., and Haslam, J. (1993). Building and using flexible models incorporating grey-level information. In: *Proc. IEEE Int. Conf. on Computer Vision* pp. 242–246, Berlin: Computer Society Press of the IEEE.

[Davatzikos & Prince, 1995] Davatzikos, C. A. and Prince, J. L. (1995). An active contour model for mapping the cortex. *IEEE Transactions on Medical Imaging,* **14** (1), 65–80.

[Kass *et al.*, 1987] Kass, M., Witkin, A., and Terzopoulos, D. (1987). Snakes: Active contour models. In: *Proc. IEEE Int. Conf. on Computer Vision* pp. 259–268, IEEE London: Computer Society Press of the IEEE.

[Kervrann & Heitz, 1994] Kervrann, C. and Heitz, F. (1994). A hierarchical statistical framework for the segmentation of deformable objects in image sequences. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp. 724–728, Seattle, WA, USA: IEEE Computer Society Press.

[Lai & Chin, 1995] Lai, K. and Chin, R. (1995). Deformable contours: Modeling and extraction. *IEEE Tran. on Pattern Analysis and Machine Intelligence,* **17** (11), 1084–1090.