A Comparative Technique and Performance Results on Novel Learned Snakes in Two Dissimilar Medical Domains

Samuel D. Fenster

Department of Computer Science
City College of New York
138th St. at Convent Ave., Rm. R8/206
New York, NY 10031

John R. Kender

Department of Computer Science
Columbia University
1214 Amsterdam, MC 0401
New York, NY 10027

Abstract

We review our work on how to teach deformable models to maximize image segmentation correctness based on user-specified criteria. We then present new variants and applications of learned snakes, modeled by four different probability density functions (PDFs), at three scales, and in the two very different medical domains of abdominal CT slices and echocardiograms.

We review and extend our method for evaluating which criteria work best. Success depends on the relation of objective function (the PDF) output to shape correctness. This relationship, for all the above learned snake variants and domains, is evaluated on perturbed ground truth shapes in three ways: by the incidence of "false positives" (scoring better than ground truth) of randomized shapes; by the monotonicity of the objective function versus shape closeness to ground truth, as given by a correlation coefficient; and by the distance of this relationship to the nearest monotonically increasing function, a new performance measure which we introduce here.

We exhaustively demonstrate such evaluations on traditional snakes, and on snakes for which image intensity and perpendicular gradient are learned separately, and with their covariances, and with separate learning over equal-length "sectors". Optimal blur appears to depend on domain. Both sectoring and the use of covariance markedly improve results in abdominal CT images, where nearby image landmarks (i.e. organs) stabilize learning. Results on echocardiograms, however, are less striking, although the use of covariance does show improvements; on investigation this appears due to the non-Gaussian distribution of image features in this domain.

1 Introduction

This work specifies and evaluates new ways of training a machine to find outlines of a known kind of object in images from a domain of consistent appearance. It uses deformable models (often called "snakes"), but this training extends the applicability of the method by adapting it where it would be unable to work otherwise. We develop novel methods to assess how well a model, trained or fixed a priori, can perform in a given domain. We assess four different models at three different resolutions in two different image domains, and find that training can provide a marked improvement, under certain conditions that our measurement techniques make precise.

There are domains in which traditional snakes, attracted to strongest or closest image edges, fail. This is true in CT images of lower abdominal organs, which are pressed up against similar organs and brighter, stronger-edged bone, and also in echocardiograms, which contain inner and outer heart walls, and large blobs of noise. This is not just a problem of avoiding suboptimal local minima—in such cases, the wrong object in fact satisfies the objective function better than the right one does. Researchers have formulated alternative objective functions to get around the problem, but without formally testing their properties in the domain, and without having some basis to compare the alternatives.

Contrary to intuition, the distribution of an objective function's values for correct contours gives no information about its goodness for segmentation, since, if incorrect contours have the same distribution of values, the function cannot guide a contour to the correct shape. We therefore examine an energy function's behavior for incorrect shapes, by generating such shapes from ground-truth shapes in a sample of domain images. Additionally, we investigate making a deformable model respond differently to local image qualities at different places on its boundary.

2 Background

2.1 Deformable models

A deformable model is a description of a shape whose parameters are iteratively adjusted until it best matches

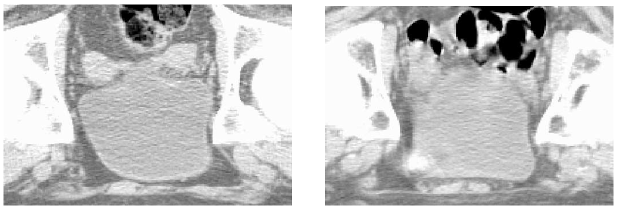

Figure 1: *In bladders of different patients, parts of their boundary vary greatly between images, like the top near intestines, but other parts are very predictable, like the sides near pelvis.*

what is depicted in an image. "Best" is measured by an arbitrary real-valued objective function, or "energy." In the case of a 1D contour in a 2D image, the model is called a snake, as in the work that first introduced the method [11].

A deformable model finds an object in an image by maximizing an objective function of image and shape: $\mathbf{S}^* = \arg\max_{\mathbf{S}} f(\mathbf{I}, \mathbf{S})$, where the $M \times N$ image $\mathbf{I} \in \Re^{MN}$ and shape $\mathbf{S}$ is a vector of shape parameters $(s_1 \ldots s_n)$. The inputs to $f\colon \Re^{MN} \times \Re^n \mapsto \Re$ are in theory, the value of every pixel and every shape parameter. The output of f reflects the likelihood that $\mathbf{S}$ depicts a particular object in image $\mathbf{I}$. Often, f is a combination of a shape energy which penalizes undesired shapes, and an image energy which responds to the strength and nearness of image edges or gradients.

To implement a deformable model, one must choose three elements. The first is a shape model, which can be a discrete set of "snaxels", or one of several continuous representations of curves or surfaces. The second is an objective function, f; a large variety of shape energies and image energies have been used. The third is an optimization technique; the usual method is some variety of gradient descent, which requires that the function be monotonic in the region searched. However, others are possible, as reviewed in [14].

One framework in which these three elements are often justified is the probabilistic formulation, which motivates the use of observed distributions, that is, learning. Here, the objective function approximates the a posteriori probability $\mathrm{P}(\mathbf{S} \mid \mathbf{I})$ that the correct shape is $\mathbf{S}$, given that the image is $\mathbf{I}$. This Bayesian formulation is: $\mathrm{P}(\mathbf{S} \mid \mathbf{I}) = \mathrm{P}(\mathbf{S})\mathrm{P}(\mathbf{I} \mid \mathbf{S})/\mathrm{P}(\mathbf{I})$ We seek the shape $\mathbf{S}^*$ most probably depicted by image $\mathbf{I}$ by maximizing $\mathrm{P}(\mathbf{S} \mid \mathbf{I})$ over $\mathbf{S}$. Noting that $\mathrm{P}(\mathbf{I})$ is constant, and $\mathrm{P}(\mathbf{I} \mid \mathbf{S})\mathrm{P}(\mathbf{S}) = \mathrm{P}(\mathbf{I} \wedge \mathbf{S})$, we maximize the simple joint probability $\mathrm{P}(\mathbf{I} \wedge \mathbf{S})$, which is a probability distribution in the single space, $\Re^{MN+n}$.

2.2 Prior work: learning for segmentation

The distributions of the image qualities must be recovered from observations in images with ground-truth contours. For deformable models, work has been done on learning specialized shape and feature models, but has not been generalized. Among learned prior shape models, there have been a multidimensional Gaussian distribution of vertex positions [7]; Markov Random Fields of vertex displacements with respect to neighbors [12, 13]; a Gaussian distribution of variations in "vibration modes" [16]; and a Gaussian distribution in a Fourier harmonic representation [18]. Among learned image features, there have been kn-dimensional Gaussians of intensities along line segments of n pixels perpendicular to a shape boundary at k feature points [7]; Gaussian distributions of multiscale intensity and gradient at a few points around many organs simultaneously [2]; 3D models of shape incorporating the observed likelihood of nearby edges being spurious [3]; histogram of pixel values from ground-truth shape boundaries [9]; and discriminant analysis on ground truth to choose among features [5].

However, in medical images some structures are neither sharp-edged nor reliably different in color from surrounding structures. Thus, despite the research, some organ contours must still be outlined manually. This is time-consuming, often rushed, and often inaccurate.

2.3 Prior work: measures of performance

Typical measures of success are based on ratios of pixels identified correctly as "in" or "out" [15]. Boundary-based measures may be more efficient, though; for example, [6] introduces a shape difference measure. Performance analysis of edge and related low-level feature detectors can be done analytically, or empirically using stochastic image noise [17], or against geometric constraints [1]. However, work on deformable model testing is sparse. In [8], a theoretical analysis shows they converge in some limited situations. But generally, when deformable models are validated experimentally, each image is tried with only one initial contour.

This work uses stochastic sampling of analytical distributions to measure performance, but randomness is applied to a different part of the system: solution shapes are randomly perturbed to determine the behavior of a candidate objective function in real images. We measure not the result of the optimization, but rather the presence of necessary qualities (local monotonicity and optimality) of the objective function.

3 Formal frameworks

3.1 Framework for training

Our inputs are a set of images for each of which the correct shape is known. Thus, the training data consists of ground truth pairs $\{(\mathbf{I}_1, \mathbf{S}_1), \ldots, (\mathbf{I}_n, \mathbf{S}_n)\}$. We must select a function that is extremal for shapes whose relation to the image data resembles those in this set. This means that a feature set $\mathbf{F}(\mathbf{I}, \mathbf{S})$, a tractable projection of $\Re^{MN+n}$, must be mapped to some measure of populatedness in the vicinity, $g(\mathbf{F}(\mathbf{I}, \mathbf{S}))$. To do this, the finite set of training

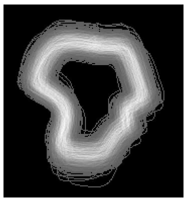

Figure 2: *1,000 perturbed versions of a heart wall contour.*

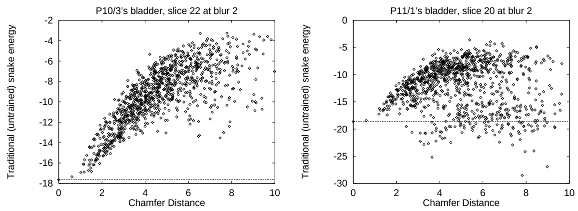

Figure 3: *Each perturbed contour is a point in a scatter plot of difference from correct shape* vs. *energy.* Left: *Plot has 0% false positives, high-correlation (0.73), is close (.55) to nearest increasing function.* Right: *Plot has 8.8% false positives, low-correlation (-0.04), is far (.98) from nearest increasing function.*

points in feature space must be extended to yield probability density values everywhere in that space; since this task is underconstrained, some kind of structure or regularization must be imposed [19]. This done, the "best" shape in an image is $\mathbf{S}^* = \arg\max_{\mathbf{S}} g(\mathbf{F}(\mathbf{I}, \mathbf{S}))$. Assuming that the training data were drawn independently, we thus must find the function g that maximizes $\prod_{i=1}^{n} g(\mathbf{F}(\mathbf{I}_i, \mathbf{S}_i))$. For Gaussian distributions, the mean and variances (and covariances, for a nondiagonal model) maximize this likelihood.

3.2 Framework for performance measurement

Performance is characterized by how the objective function reacts to near-correct shapes in actual, complex image data. An objective function must satisfy two kinds of optimality: absolutely, the ground-truth shape must score better than any other shape, and (for gradient-based algorithms) relatively, as a shape approaches ground truth its score must improve. Thus, the basis of our test of an objective function will be the generation of m shape samples per image (in practice, m=1,000), each sample determining the pair (D, f), where D is a measure of how close the sample is to the ground truth shape, and f is the objective function value for the sample.

Several well-known measures of shape difference exist. The Hausdorff distance [10] would not be a good indicator here, as it treats a single, local n-pixel deviation the same as displacement by n pixels everywhere. The symmetric difference of the areas enclosed by two boundaries [15] is inefficient to calculate. An elegant iterative measure is in [6]. But we use the chamfer distance between two boundaries [4], the average over one shape of distance to the closest point on the other. Although this measure is asymmetric, we found in our imagery that it has a high correlation (0.977) between $D(a, b)$ and $D(b, a)$. And although arbitrarily long but narrow protrusions may not register large differences, these did not arise in our domains.

For each test image, the relationship of objective function to known shape correctness (distance to ground truth) over the expected range of shape samples is described exactly by a scatter plot. We use three statistics to summarize scatter plot properties. The most straightforward test of absolute optimality is the number of contours having energies below that of the ground truth, falsely indicating that they are better solutions. An approximate test for relative optimality (peakedness) uses a plot's correlation coefficient to measure monotonicity. But a third and ideal indicator of robust gradient descent behavior would be how far the point set $\{(D_i, f_i) : 1 \leq i \leq n\}$ is from the closest increasing function. Thus, the root-mean-square residual distance, normalized by the variance of the f_i, is a measure of monotonicity which varies from 0, if the data is strictly increasing ("good"), to 1, if it is strictly decreasing ("bad"). Randomly fluctuating data, with 1,000 samples, has been found to yield measures within 1% of 1 (very "bad").

4 Implementations

4.1 Implementing training

Shapes are currently limited to piecewise cubics; those we have used are polylines and C^2 cubic splines. This design limits objective functions to those which are sums over the points in the shape and its normals, but it makes f independent of the shape model $\mathbf{S}(u)$. 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 $\mathbf{F}(\mathbf{I}, \mathbf{S})$. Blurring scales of $s = 2$, 4 and 8 pixels were tried, based on observation of image properties and structures.

We chose perhaps the simplest model of the distribution of our features—a multidimensional Gaussian. Specifically, we assumed independent, identically-distributed values of intensity and directional gradient 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.,* $\mathrm{P}(\mathbf{F}(\mathbf{I}, \mathbf{S}))$. This is a product of Gaussians; its negative log, "image energy"

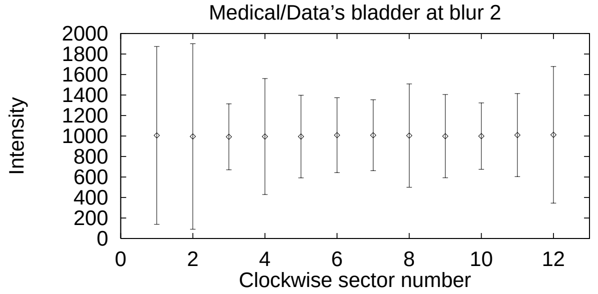

Figure 4: *Each sector has separately learned a Gaussian of image intensity. The range bars are tallest in regions 12, 1 and 2, in the greatest intensity variability at the top of the bladder.*

E, is E_A, defined as:

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

We also used an energy with one more parameter, the correlation between intensity and gradient at a point. This is a 2D Gaussian rather than the joint distribution of two 1D Gaussians. This energy was defined as

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

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

To provide comparison with traditional methods as applied in the same domain, we also used a "traditional" objective function (not based on training) that summed gradient strengths on the shape boundary:

$$E_T = -\oint ||\nabla \mathbf{I}_s(\mathbf{S}(u))|| du$$

Having noted that the qualities that the model should seek may not be uniform everywhere on its boundary, we also developed a spatially varying objective function, which we call sectored snakes. In this, we followed [7], although they trained a small number of preselected feature points. We seek an accurate boundary everywhere, so we divided the contour into a fixed number of equal-length regions (sectors), each with separate training (Figure 4). (We chose 12 sectors, corresponding to clock positions about the center of mass of the contour; this was based on the sizes of surrounding structures in our imagery.)

4.2 Implementing performance measurement

The novel measure of closeness to nearest increasing functions was programmed from basic principles, and is described here. Assume function $f(x)$ and data $S = \{(x_i, y_i) : 1 \le i \le n\}$. We define the "distance" of S from f as the sum of squared differences of residuals, $\sum (y_i - f(x_i))^2$. If one increasing (continuous) function f_0 minimizes this distance, then so do many others, since the distance only depends on f's values at the x_i. Thus, the problem is reduced to finding an increasing (discrete) set of $f(x_i)$, or f_i, that minimizes the sum of residuals; further, this minimum does not depend on the values of the x_i, but only on the ordering they impose on the f_i. Since it is the residual error which gives a measure of how close the data set is to an increasing function, an algorithm to find the error therefore only needs as input an ordered sequence of y_i.

The algorithm for finding $\{f_i\}$ is based on the insight that the closest increasing sequence to a decreasing sequence of y_i is the constant sequence consisting of its mean. An initial sequence $\{f_i^0\}$, not yet made increasing, is set equal to the data. In successive iterations $\{f_i^k\}$, decreasing subsequences are selected and then replaced by equal length subsequences consisting of that many repetitions of the subsequence mean. These constant runs can be stored efficiently; the algorithm can be proved to terminate; and the resulting sequence, modified in place, is the closest (in terms of the least-squares, L^2, norm) nondecreasing sequence to the input sequence.

The shape perturbations were a combination of translation and of scaling independently around two randomly chosen orthogonal axes centered on the contour's centroid. The translations, and the logs of the scale factors, were normally distributed. Since the objects were inner heart walls with widths of 40–120 pixels and heights of 30–90 pixels, and bladders, with widths of 70–90 and heights of 70–100, we considered it reasonable that the translations were given a standard deviation of 5 pixels, and that the logs of the scale factors were given a standard deviation of $\log(1.1)$, or 10% stretch/shrinkage.

In each of our two domains, we tested twelve different objective functions. There were four different feature models—the traditional untrained snake; one using Gaussians modeling intensity and directional gradient; the same with covariance; and the same with 12 independently modeled sectors. Each was tested at three scales, blurring with Gaussian kernels of σ=2, 4 and 8 pixels. All tests used the same 1,000 parameterized perturbations.

5 Experiments

To the best of our knowledge, this is the most extensive formal measurement and comparison of the effectiveness of deformable model energy functions.

5.1 Domains

In the domain of bladders in abdominal CT scans, our images were from a study at Memorial Sloan-Kettering Cancer Center. Upon investigation, it appeared that the

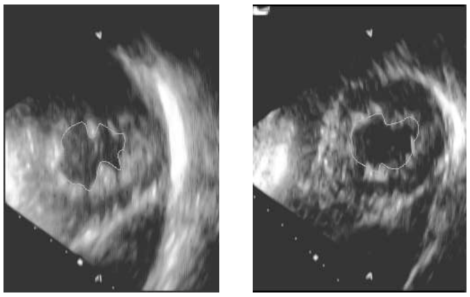

Figure 5: *Two representative echocardiograms, showing degree of speckle and inexactness of ground-truth contour.*

"ground truth" available to us was not drawn very accurately; it is often about six pixels away from the visible organ boundary, in a 512×512 image. 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 under time pressure. We trained and tested on 36 high-quality bladder contours.

In the domain of ultrasound images of the heart (echocardiograms), the quality of the images is much poorer. Regions are not of homogeneous color but are rather characterized by "speckle" concentrated at structure boundaries, especially those perpendicular to the sensor direction (Figure 5). We trained and tested on the rather inexact expert-drawn ground truth boundaries of the inner heart wall, taken from 24 patients, at end-systoly (maximum contraction) and on the short-axis views. The images we used, from the commercial firm, Echovision, had been preprocessed using their proprietary (and therefore unknown) despeckling algorithm.

5.2 Protocol

Training and segmentation was done for four different snake formulations: the sectored snake; the simple trained model; the model with covariance; and a traditional, untrained snake whose energy is gradient strength traversed by the contour. Each of these was tested at three scales, $s = 2$, $s = 4$, and $s = 8$, for a total of twelve different objective functions. We evaluated each of these energy functions with 1,000 perturbations of the ground truth contour. We used "jackknife testing": the performance in each ground-truth image was tested using models trained on all images but that one. We evaluated each with our three measures (false positives, correlation, distance to closest increasing function), and characterized each function's performance in each domain by giving each statistic's average, average deviation, and confidence interval over the entire ground truth test set. Average deviation is similar to the standard deviation, but is more robust to outliers. The confidence intervals are, however, based on standard deviation, and are therefore worst-case.

5.3 Bladder Results

Overall, training provided a marked improvement over simple untrained attraction to edges.

The simplest form of training using two independent Gaussians, modeling the probability of finding a given intensity $\mathbf{I}_s(\mathbf{S}(u))$ at a contour pixel with blur s, and of finding a given gradient $\mathbf{S}^{\perp}(u) \cdot \nabla \mathbf{I}_s(\mathbf{S}(u))$ perpendicular to the contour. By way of illustration, Table 1 shows the distribution parameters μ_I, σ_I and μ_∇, σ_∇ of these features at the best scale, $s = 2$. It also shows $\rho_{I\nabla}$, the additional parameter used by the model that learns the two Gaussians with covariance. The Gaussian model provided a reasonable fit; see Figure 6.

The variance of the gradient is high at coarse scales; this would seem to make the gradient an insignificant contributor to the objective function. Nevertheless, the function that modeled the covariance of the intensity and the gradient produced half the false positives of its closest competitor at that scale, meaning that useful information was in fact present in the gradient's high correlation with intensity, $\rho_{I\nabla}$, despite the gradient's high variance on its own.

False positives (see Table 2) occurred far less often with trained snakes than traditional snakes—in fact, the incidence was negligible when the least blur was used, for all three varieties of training. By contrast, a traditional snake, attracted to the strongest edges, incorrectly gave 15% of the perturbed shapes a better score; at higher blurs, it did even worse. At coarser scales, differences among the kinds of training became apparent. At scale 4, sectoring cut errors almost in half, as did use of covariance.

The correlation of distance from ground truth *vs.* objective function is in Table 3. Most importantly, all of the trained formulations demonstrated a much higher correlation than traditional untrained snakes. Snakes with sectored training had slightly higher correlations than unsectored trained snakes at bigger image scales. For all objective function models, coarser (blurrier) image scale improved correlation, but coarser scale also devastated false-positive rates.

As hoped, a very nearly inverse relationship was observed between the correlation coefficient and our novel measure of distance from monotonicity; thus, in this domain, both are equally good indicators. As a comparison, the value of this measure for randomized scatter plots were experimentally computed to be uniformly between 0.99 and 1; objective functions with significant monotonicity in Table 4 have much lower values.

5.4 Heart Results

Results were generally poorer because of the quality of the images and contours; still, many useful relationships emerged.

Table 5 gives an illustration of parameters recovered in

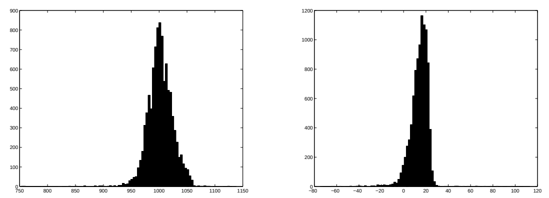

Figure 6: *Histograms of pixel intensity (left) and perpendicular gradient distribution for bladders (blur scale 2).*

Scale	μ_I	σ_I	μ_∇	σ_∇	$\rho_{I\nabla}$
2	1002	23	14	8.9	-0.55
4	1004	26	5.9	7.1	-0.77
8	1010	31	1.6	4.1	-0.86

Table 1: *Bladder CT: learned parameters (9,700 pixels in 36 contours of 24 image sets).*

Objective function	Scale (pixels)	False positive rate Avg %	Avg dev	95% conf
Untrained (traditional)	2	15	15	±5.9
	4	23	17	±6.3
	8	32	20	±7.7
Single intensity & gradient strength Gauss	2	1.1	1.6	±0.98
	4	11	11	±4.1
	8	33	5.8	±2.7
Gaussians in 12 sectors	2	1.1	1.7	±1.1
	4	6.7	7.4	±3.3
	8	27	6.4	±3.0
Gaussians with covariance	2	1.6	2.5	±1.4
	4	6.6	7.9	±3.0
	8	14	12	±4.4

Table 2: *Performance on bladder CT: false positives (1,000 perturbed contours for each of 36 images, using jack-knife testing).*

Objective function	Scale (pixels)	Correlation coefficient Avg	Avg dev	95% conf
Untrained (traditional)	2	0.15	0.34	±0.12
	4	0.21	0.37	±0.14
	8	0.22	0.38	±0.15
Single intensity & gradient strength Gauss	2	0.55	0.14	±0.057
	4	0.58	0.14	±0.056
	8	0.44	0.089	±0.040
Gaussians in 12 sectors	2	0.55	0.12	±0.050
	4	0.60	0.14	±0.055
	8	0.51	0.083	±0.038
Gaussians with covariance	2	0.54	0.095	±0.037
	4	0.60	0.11	±0.039
	8	0.63	0.12	±0.045

Table 3: *Performance on bladder CT: correlation coefficient, between chamfer distance and image energy.*

Objective function	Scale (pixels)	Obj. function monotonicity Avg	Avg dev	95% conf
Untrained (traditional)	2	0.87	0.16	±0.058
	4	0.84	0.20	±0.075
	8	0.85	0.18	±0.070
Single intensity & gradient strength Gauss	2	0.76	0.13	±0.051
	4	0.73	0.14	±0.052
	8	0.81	0.099	±0.046
Gaussians in 12 sectors	2	0.76	0.12	±0.044
	4	0.71	0.14	±0.054
	8	0.74	0.10	±0.052
Gaussians with covariance	2	0.79	0.079	±0.032
	4	0.73	0.098	±0.037
	8	0.68	0.13	±0.049

Table 4: *Performance on bladder CT: monotonicity (closeness to nearest increasing function).*

the simple two-Gaussian training. Variance in both features was large, meaning that the probability distribution induced by the data was estimated with poor certainty. Correlation between the two features, though, was significant and useful, as further testing showed. The actual separate distributions which the Gaussians attempt to model are histogrammed in Figure 7.

False positive rates were unacceptable for all functions tested. At the finest scale, untrained snakes, seeking strongest edges, did somewhat better than any of the trained models, showing that one objective function does not fit all possible situations. One of the likely reasons that Gaussian training could not recognize correct boundaries at this scale well is that both intensity and gradient along the desired boundaries in any given image appear to be bimodal. (Figure 7 appears unimodal because it is an aggregate of features in 24 images.) Inspection of the images shows that speckle makes intensities along the boundary alternate from dark to light, and gradients perpendicular to the boundary alternate in polarity.

However, at coarser scales, the untrained snake's performance was eclipsed by the trained snakes, particularly the one that modeled covariance. The covariant model worked best at a higher blur (4) here than it did in bladder CT imagery, due to the effects of blurring on speckle.

Again, as measured by correlation coefficient or by distance to nearest increasing function, the untrained snake scored best of all the models—at the finest scale. But unlike in the CT domain, increasing the blur made the untrained model perform significantly worse, not better. With the Gaussian trained models, blur helped. At high blur, modeling covariance provided significant cues as to distance from shape correctness. Not too surprisingly, sectoring provided no advantage in this speckled rather than cluttered domain. This indicates (as is visually apparent) that in this domain image qualities did not differ consistently according to what portion of the contour they were on, or according to what stabilizing organs they were near; in fact, there were few localizing features of any kind, and no stabilizing bony structures at all.

Scale	μ_I	σ_I	μ_∇	σ_∇	$\rho_{I\nabla}$
2	869	428	-52	75	-0.48
4	823	372	-32	42	-0.53
8	764	324	-14	19	-0.48

Table 5: *Echocardiogram: learned parameters (7,073 pixels in 24 contours in 24 image sets).*

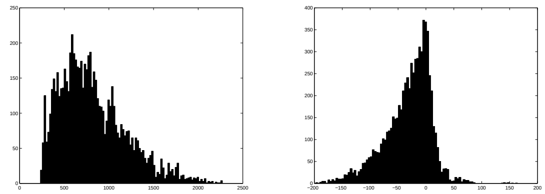

Figure 7: *Histograms of pixel intensity (left) and perpendicular gradient distribution for inner heart walls (blur scale 4).*

Objective function	Scale (pixels)	False positive rate Avg %	Avg dev	95% conf
Untrained (traditional)	2	18	15	±7.0
	4	32	21	±10.0
	8	47	18	±8.4
Single intensity & gradient strength Gauss	2	31	21	±10.0
	4	29	20	±9.2
	8	34	18	±8.5
Gaussians in 12 sectors	2	29	21	±10.0
	4	28	21	±9.6
	8	35	18	±9.0
Gaussians with covariance	2	27	17	±8.6
	4	23	16	±8.0
	8	30	16	±7.9

Table 6: *Performance on echocardiograms: false positives.*

Objective function	Scale (pixels)	Correlation coefficient Avg	Avg dev	95% conf
Untrained (traditional)	2	0.55	0.15	±0.071
	4	0.38	0.23	±0.12
	8	0.11	0.31	±0.15
Single intensity & gradient strength Gauss	2	0.23	0.26	±0.13
	4	0.34	0.26	±0.13
	8	0.35	0.23	±0.11
Gaussians in 12 sectors	2	0.24	0.25	±0.13
	4	0.33	0.27	±0.13
	8	0.33	0.25	±0.12
Gaussians with covariance	2	0.19	0.23	±0.12
	4	0.34	0.19	±0.11
	8	0.40	0.20	±0.11

Table 7: *Performance on echocardiograms: correlation.*

Objective function	Scale (pixels)	Obj. function monotonicity Avg	Avg dev	95% conf
Untrained (traditional)	2	0.78	0.10	±0.049
	4	0.85	0.12	±0.063
	8	0.93	0.078	±0.042
Single intensity & gradient strength Gauss	2	0.90	0.072	±0.033
	4	0.85	0.11	±0.052
	8	0.82	0.13	±0.064
Gaussians in 12 sectors	2	0.90	0.078	±0.039
	4	0.85	0.12	±0.057
	8	0.83	0.13	±0.060
Gaussians with covariance	2	0.93	0.054	±0.029
	4	0.87	0.064	±0.032
	8	0.83	0.10	±0.052

Table 8: *Performance on echocardiograms: monotonicity.*

6 Conclusions

In our domains of cluttered abdominal CT imagery, the enhancements we made to snakes were necessary to get good segmentation results. Our evaluation tools provide evidence that heterogeneous training (sectoring) reduces the learned objective function's false positive rate significantly over simple unsectored training, particularly where there was systematic variation around the object being outlined.

The CT evaluations show much better performance for energies based on the fine image scale than for those on the coarse scale; the ultrasound evaluations show the opposite. This demonstrates that domain-dependent testing is necessary to measure and choose the best objective function. Particularly in domains in which object boundaries were less accurately drawn, a coarser scale makes image properties on the boundary "visible" to the contours that are further away. 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 particularly in imagery such as these, gradual deblurring is indicated—a standard robust optimization technique recommended in [11].

In summary, we have demonstrated and evaluated straightforward training of a continuous shape model. We have explored sectoring, a modification of snakes that incorporates local adaptability, and that gives quantifiable improvement at only a small increment in complexity. We have developed and demonstrated a methodology that allows us (and others) to quantify how good deformable models are, and to gain insight into an objective function's strengths. As part of this method, we have contributed a new way of measuring a data set's proximity to an increasing function that is more justifiable than simple correlation. And lastly, we have tested the methodology on a believably large data set of images in two very different domains.

Acknowledgements

We would like to thank Drs. Radhe Mohan and Chen Chui of the Medical Physics Computer Services Department at Memorial Sloan-Kettering Cancer Center, and Dr. Jeffrey Weisman of EchoVision in Philadelphia, for their expertise and their image and contour data.

References

[1] Simon Baker. *Design and Evaluation of Feature Detectors*. PhD thesis, Columbia University, 1998.

[2] Bernard Baldwin. *Multiscale Snakes*. PhD thesis, Courant Institute of New York University, 1997.

[3] Jennifer L. Boes, Charles R. Meyer, and Terry E. Weymouth. Liver definition in CT using a population-based shape model. In *Proc. 1st Int'l Conf. on Computer Vision, Virtual Reality, and Robotics in Medicine (CVRMed '95)*, pages 506–512, Nice, France, April 1995. Springer.

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

[5] M. J. Byrne and J. Graham. Application of model based image interpretation methods to diabetic neuropathy. In *Proc. European Conf. on Computer Vision*, volume II, pages 272–282, Cambridge, UK, April 1996. Springer.

[6] Vikram Chalana and Yongmin Kim. Methodology for evaluation of boundary detection algorithms on medical images. *IEEE Tran. on Medical Imaging*, 16(5):642–652, October 1997.

[7] T.F. Cootes, A. Hill, C.J. Taylor, and J. Haslam. Building and using flexible models incorporating grey-level information. In *Proc. IEEE Int'l Conf. on Computer Vision*, pages 242–246, Berlin, May 1993. IEEE Computer Society.

[8] C. Davatzikos and J.L. Prince. Global minimum for active contour models: A minimal path approach. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 674–679, San Francisco CA USA, June 1996. IEEE Computer Society.

[9] Robert P. Grzeszczuk and David N. Levin. "Brownian strings": Segmenting images with stochastically deformable contours. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 19(10):1100–1114, October 1997.

[10] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the Hausdorff distance. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 15(9):850–863, September 1993.

[11] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. In *Proc. IEEE Int'l Conf. on Computer Vision*, pages 259–268, London, 1987. IEEE Computer Society.

[12] Charles Kervrann and Fabrice Heitz. A hierarchical statistical framework for the segmentation of deformable objects in image sequences. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 724–728, Seattle, WA, USA, June 1994. IEEE Computer Society.

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

[14] F. Leymarie and M.D. Levine. Tracking deformable objects in the plane using an active contour model. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 15(6):617–633, June 1993.

[15] P. P. Ohanian and R. C. Dubes. Performance evaluation for four classes of natural textures. *Pattern Recognition*, 25(8):819–833, August 1992.

[16] A. P. Pentland and S. Sclaroff. Modal matching for correspondence and recognition. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 17(6):545–561, June 1995.

[17] Visvanathan Ramesh. *Performance Evaluation of Image Understanding Algorithms*. PhD thesis, University of Washington, 1995.

[18] Lawrence H. Staib and James S. Duncan. Deformable Fourier models for surface finding in 3D images. In *Visualization in Biomedical Computing*, volume 1808, pages 90–104, Chapel Hill, NC, USA, October

1992. SPIE–the Int'l Society for Optical Engineering, SPIE.

[19] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, NY, USA, 1995.