Nonparametric Training of Snakes to Find Indistinct Boundaries

Samuel D. Fenster
Computer Science Dept.
City College of New York
New York, NY 10031
fenster@cs.ccny.cuny.edu

Chun-Bin Gary Kuo
Computer Science Dept.
Columbia University
New York, NY 10027

John R. Kender
Computer Science Dept.
Columbia University
New York, NY 10027
kender@cs.columbia.edu

Abstract

We enable highly improved performance of deformable model (snake) segmentation of a known type of object (human bladder) with unclear edges in a cluttered domain (abdominal CT scans). This is accomplished by learning an objective function from ground-truth contours in test images, using a nonparametric estimator of the distributions of chosen image quantities (intensity on the boundary and image gradient perpendicular to it). The Parzen-window estimator is found to reward correct contours much more accurately than a model based on means and covariances. This latter Gaussian model, in turn, performs adequately where a traditional a priori *objective function does not. Performance of objective functions is measured by checking the fraction of incorrect contours that score better than ground truth (false positives), and the deviation of plots of shape incorrectness* vs. *objective function value from the closest strictly increasing function.*

1. The framework

It can be difficult to find the boundaries of a structure in a complex image. There may be nearby edges that are sharper than those of the object sought; it may be unclear what point in the transition from light to dark is considered the boundary, for purposes of a given task in a given image domain; the boundary may be best characterized by some feature other than image gradient, such as texture. In such cases, learning from images with known ground-truth boundaries can supply parameters that define where a boundary most likely is. If the criterion that best predicts the correct separation of pixel regions is unknown, then ground truth can be used to test the effectiveness of segmentation algorithms based on different criteria. We report here on the success of a Parzen-window model used for segmentation based on intensity at the boundary and on gradient perpendicular to it. We compare this to a somewhat poorer-performing Gaussian model of the same features, and to a much worse traditional criterion. The domain is abdominal CT scans used for radiation treatment of prostate cancer; for our tests, we sought the boundary of the bladder.

The framework of criteria with parameters lends itself well to implementation as objective functions, and the method we employ for this study, the deformable model [1], provides a way to use objective functions for image segmentation. In it, a parametric model of boundary shape is set to an initial guess for an image. An optimization method (often gradient descent) then adjusts its parameters to maximize an objective function of image and shape. The traditional objective function simply sums edge strengths on the shape boundary. But we believe the method is particularly useful in tasks and image domains for which a different objective function is needed. Cluttered medical images have proven to be such a domain (Figure 1).

The performance of different objective functions in an image domain can be compared: In a sampling of test images, objective functions values are observed for shapes randomly perturbed by different amounts from the ground truth [2]. For each objective function, this allows us to quantify how close to "optimal" it rates the ground truth; and, if gradient descent optimization is used, we can measure the degree to which the objective function gives worse scores to shapes that are further from ground truth.

2. Training a deformable model

Any deformable model segmentation algorithm uses a parametrized shape model, $\mathbf{s}(u;\mathbf{S}) \mapsto \Re^2$ in 2D. Given an image $\mathbf{I}$ and an initial guess at the shape (parameter vector $\mathbf{S}$), it alters (deforms) the shape so as to maximize an objective function of shape and image. (This is often done iteratively by gradient descent, which actually only uses the derivative of the objective function with repect to shape parameters.)

A typical traditional objective function f sums image gradient magnitude $||\nabla \mathbf{I}(\mathbf{x})||$ over the points on a shape boundary: $f(\mathbf{I},\mathbf{S}) = \sum_u ||\nabla \mathbf{I}(\mathbf{s}(u;\mathbf{S}))||$. But when such a function does not suffice, a function based on observed

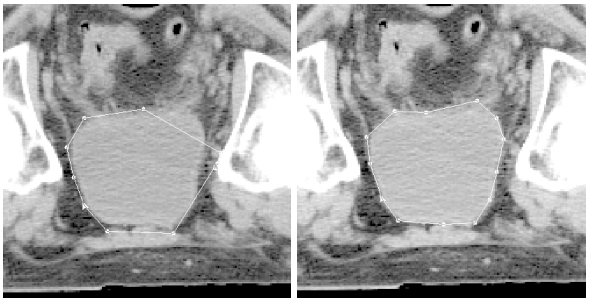


Figure 1: A CT scan of the bladder. A traditional objective function, summing edge strength around a candidate contour, scores the contour on the left higher than the one on the right. An objective function trained with ground truth from the task domain gives the one on the right a higher score.

statistics on qualities of the desired boundary may succeed.

To find the boundary that has *maximum likelihood* (ML) based on some observable quantities $\mathbf{F}(\mathbf{I}, \mathbf{S})$ in an image, the probability distribution $g(\mathbf{F})$ of those quantities must be known. Then the shape $\mathbf{S}$ optimizing $g(\mathbf{F}(\mathbf{I}, \mathbf{S}))$ can be found. For example, the features we used were σ-blurred intensity on the boundary, $\mathbf{I}_\sigma(\mathbf{s}(u; \mathbf{S}))$, and image gradient perpendicular to it, $\mathbf{s}^\perp(u; \mathbf{S}) \cdot \nabla \mathbf{I}_\sigma(\mathbf{s}(u; \mathbf{S}))$, observed at arc-length intervals of one pixel. (Others have often incorporated image-independent shape qualities.) We tried modeling the distribution of these quantities as a joint Gaussian distribution, and then as a Parzen-window distribution. For both, we assumed that every contour pixel was identically distributed (we have tested spatially-varying distributions elsewhere [2]); and that pixels were independent, and thus the probability of a full set of intensities or gradients around a curve is the product of probabilities $g_p(x)$ of feature values x at each point. For example, for intensity:

$$f(\mathbf{I}, \mathbf{S}) = g(\mathbf{F}(\mathbf{I}, \mathbf{S})) = \prod_u g_p(\mathbf{I}_\sigma(\mathbf{s}(u; \mathbf{S})))$$

A probability distribution is recovered from observations of the chosen features $\mathbf{F}$ in a set of training images $\{\mathbf{I}_1, \ldots, \mathbf{I}_n\}$ in which correct shape boundaries $\{\mathbf{S}_1, \ldots, \mathbf{S}_n\}$, the ground truth, are known. It is a member of a family $\mathcal{G}$ of probability density functions (PDFs); the family is a *probability model*. The member $g \in \mathcal{G}$ which gives the highest joint probability for observing the feature values actually seen in the training set is the maximum-likelihood PDF. If training images were randomly drawn, then we choose the PDF g which maximizes $\prod_i g(\mathbf{F}(\mathbf{I}_i, \mathbf{S}_i))$.

The family $\mathcal{G}$ from which g is drawn determines the form of the PDF function g. If it is a family of multidimensional Gaussians, then the maximum-likelihood PDF is that with the means, variances and covariances of feature-values that are observed in the training set.

3. Parzen-window probability model

Another probability model is one which will take the shape of the training data, no matter what that shape is, rather than being confined to a family of distributions with a small number of degrees of freedom. The distribution of feature values (*e.g.*, intensities on the boundary) in the training data itself is a sum of spikes, or delta functions, one for each observed occurence of a value. This would be a truly "nonparametric" distribution derived from training. Such methods have been used before, although without a comparative performance analysis: for example, Grzeszczuk and Levin [4] trained a snake based on a 2D histogram of intensity inside *vs.* outside a snake boundary.

But training data is only useful if the resulting distribution gives a nonzero probability to a feature value that is close, but not identical, to one observed in training. So an occurence of a value in training must be taken as evidence that nearby values are also likely. So each observed value can contribute some kernel, centered around it, to the PDF. This is known as a *Parzen-window* estimator [3]. If the probabilities of values near the observed one are taken to be Gaussian, G_s, then the standard deviation s determines how probable is a value different by one unit, based on the evidence of the observed value. As more identical or similar values appear in a cluster, the value of the (smooth) PDF in that region grows. Though the Parzen approach is called "nonparametric," it still requires a choice of kernel and kernel width.

Thus, if N feature value observations (for instance, intensities) $x_1, \ldots, x_N$ were made during training, then the estimated likelihood of a feature value x, observed in a candidate shape in an image, is:

$$g(x) = \frac{1}{N} \sum_{k=1}^{N} G_s(x - x_k)$$

We tried a trained Parzen-window objective function based on a Gaussian kernel. Since a Parzen estimation of a multidimensional (joint) PDF would require a combinatorially large array of terms, we made the simplifying assumption that the multiple quantities observed in one image (intensities on the boundary and perpendicular gradients) were independently distributed, allowing the multiplication of two one-dimensional PDFs. By contrast, a multidimensional Gaussian PDF requires only a small array of covariances. We tested a Gaussian model that also assumed independence, and one that had covariances.

It is not clear how to test whether evidence contributed by a training observation is actually Gaussian around that observation; so the choice of a Gaussian kernel, though reasonable, is not derived from analysis or measurement.

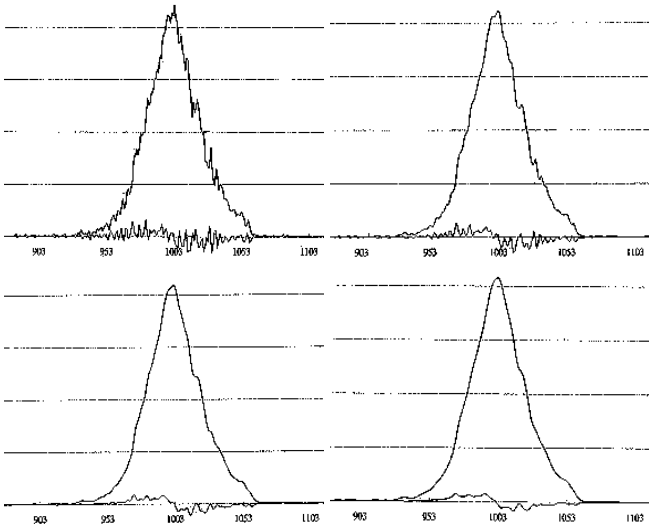


Figure 2: Parzen estimator of intensity PDF (high curve) and its derivative (low curve) with Gaussian windows of σ =0.5, 1.0, 1.5 and 2.0. Only 1.5 and above eliminates the numerous local extrema, zero crossings in the derivative, which will attract the deformable shape to incorrect intensity values.

The width of the kernel was chosen as a standard deviation of two intensity units, the smallest that yielded a smooth PDF rather than one characterized by small spikes (see Figure 2). Thus, this choice was not strictly computational: humans decided at what scale the data displayed modes rather than separate spikes from individual values. The data for our domain appeared unimodal. In CT data, soft tissue intensity was around 1000, and air was 0. For real-valued perpendicular gradient values, data was accumulated in bins of size 2.

4. Measuring objective function performance

To assess the effectiveness of a deformable model, we needed a way to characterize an objective function's performance, whether learned or specified *a priori*.

We work in negative logs of probabilities. Therefore, if a deformable model's objective function yields a smaller value in every image for the ground truth contour than for any other contour nearby, then an ideal optimization process will guarantee a correct segmentation. If not, it must occasionally fail. We can estimate a function's closeness to this guarantee by generating perturbations of ground truth in representative test images, and seeing what fraction of them generate objective function values smaller than the ground truth. Ideally, none of them will. We call this measure the "false positive" rate. The perturbed shapes only need be as far from ground truth as the initial shape guess, and intermediate shapes explored by optimization, are expected to be. But perturbations should not be generated that are within tolerance for correctness of the supplied ground truth. We generate 1000 random perturbations using normally distributed translations with standard deviation of 5 pixels, and scaling with a standard deviation of 10% independently along two random orthogonal axes. The distribution created for one shape is shown in Figure 3, left.

If optimization is by gradient descent, then convergence to the correct shape is likely if shapes that are closer to correct always have lower objective function values (the negative gradient in shape parameter space points toward them). So assume a measure of shape distance. Then a plot of distance from ground truth *vs.* objective function value, for perturbed contours, should monotonically increase. Figure 3 shows 1000 random perturbations of a ground-truth bladder contour, and two resulting scatter plots of shape distance *vs.* objective function value.

In practice, no function will strictly increase without exception; and even a theoretically perfect function only needs to increase along each single path of deformation, not for any two widely separated perturbations. But a real-valued (rather than Boolean) measure of closeness to an increasing function can provide an indicator of the likelihood of local objective function minima other than the ground truth shape. We use the RMS distance from the data to the nearest increasing sequence. Since y-axis units should not affect a "monotonicity" measure, we normalize by variance, resulting in a measure that ranges from 0 (data is strictly increasing) to 1 (data is strictly decreasing). Random data tends to have a measure of 0.99, almost as far as possible from the nearest increasing sequence.

The shape distance measure used is chamfer distance [5]. The nearest increasing sequence is the complete-ordering case of a well-known statistical technique, isotonic regression. The efficient Pool Adjacent Violators algorithm of Ayer, Brunk *et al.* [6], among others, finds the closest increasing function to the perturbation test data.

5. Results

Thirty-six images with known ground truth were tested. Each was tested using objective functions leaned from the other 35. Below we see the "false-positive" rate and monotonicity for a traditional sum-of-edge-strength objective function, and for functions based on likelihood of intensity and directional gradient on the contour, as learned with several different probability models — independent Gaussians for intensity and directional gradient; a 2D Gaussian with covariance; and independent Parzen-window models. All assume identical distributions for every pixel on the contour. They were tested on unblurred images, and on images

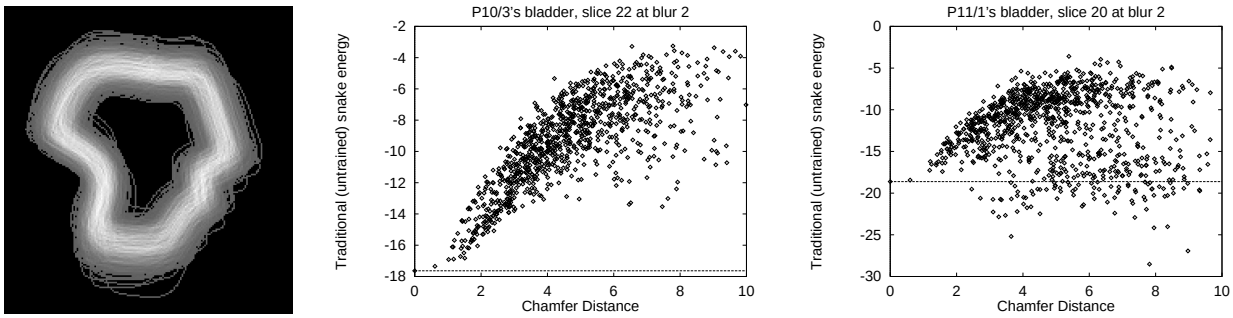
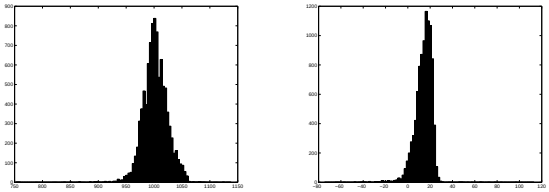


Figure 3: *Left:* 1000 perturbed versions of a heart wall contour. Each perturbed contour is a point in a scatter plot of difference from correct shape *vs.* objective function value. *Middle:* Plot has 0% false positives (below the dotted line), and is close (.55) to the nearest increasing function. *Right:* Plot has 8.8% false positives, and is far (.98, with the worst possible being 1.0) from the nearest increasing function.

convolved with Gaussians of standard deviation 2, 4 and 8.

Figure 4 shows histograms and statistics of the distributions of the intensities and perpendicular gradient values found around all 36 contours in the ground-truth test set. They are unimodal, but, as the Kolmogorov-Smirnov test results in Table 1 show, with high confidence, they are not Gaussian.



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

Figure 4: Bladder CT: Distribution and learned parameters. Distributions were not Gaussian, but similar-shaped Gaussian model may succeed.

# of pixels	K-S value	K-S min for confidence of 90%	95%	99%
9700	0.043	0.008	0.009	0.010

Table 1: The Kolmogorov-Smirnov test simply returns the maximum difference between two cumulative distributions. Here we test our bladder contour intensity data against Gaussians with the same mean and variance. Although the distribution of intensities is bell-shaped, its K-S value exceeds that for which there is 99% confidence that it is *not* Gaussian.

Table 2 compares the false positive rate for objective functions based on the different estimators and image blurs, in test images from our domain. (Comparative performance might differ greatly in some other domain.) We see that the traditional, untrained snake has a high false positive rate, reflecting cases like that shown in Figure 1.

The uncorrelated Gaussian estimators for intensity and perpendicular gradient on the boundary provide an objective function which, on an unblurred image, score less than 1% of perturbed contours better than unperturbed ground truth. The correlated Gaussians perform the same at their best, but twice as well for higher blurs where the performance was worse.

The Parzen estimator, though, has one-third to one-fifth of the false positives of the *correlated* Gaussians. The best performance is with no image blur, where the false positive rate is less than 1 in 700. This improvement over the Gaussian models is probably because of skew in one of the feature distributions, causing the Gaussian's maximum to be at the wrong feature value.

Another condition for good segmentation performance, when gradient descent optimization is used, is that an objective function be close to a monotonically increasing function with respect to the shape's distance from ground truth. Once again, as Table 3 shows, the untrained model performs the worst. With no blur, the other models do no better: A blurred image is best at guiding a distant shape toward ground truth, even though an unblurred image best allows it to settle in the correct spot.

But in blurred images, the Gaussian models have significantly better values of this measure, and the Parzen-window function, in turn, does much better than the Gaussians.

Performance on Bladder CT: False Positives

Objective function model	blur (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
Joint intensity & gradient strength Gaussians	0	**0.79**	1.3	±0.99
	2	**1.1**	1.6	±0.98
	4	**11**	11	±4.1
	8	**33**	5.8	±2.7
2D Gaussian with covariance	0	**0.79**	1.3	±0.94
	2	**1.6**	2.5	±1.4
	4	**6.6**	7.9	±3.0
	8	**14**	12	±4.4
Joint Parzen model	0	**0.14**	0.16	±0.11
	2	**0.18**	0.20	±0.10
	4	**1.5**	1.9	±0.96

Table 2: False-positives in the bladder CT image domain, for a traditional snake objective function and three probability models of image intensity on the boundary and perpendicular image gradient, with various Gaussian image blurs. The unacceptable traditional objective function always produced at least 15% false positives. Gaussian models gave a rate of less than 1%, and the nonparametric Parzen-window PDFs gave a rate of one-fifth that.

Average deviation of false positives among the 36 test images is like standard deviation, but less sensitive to outliers. The false positive rate is within the stated limits with 95% confidence, assuming Gaussian sampling error.

Performance on Bladder CT: Monotonicity

Objective function model	blur (pixels)	Obj. function monotonicity avg	avg dev	95% conf
Untrained (traditional)	2	**.86**	.16	±.060
	4	**.83**	.21	±.078
	8	**.84**	.18	±.068
Joint intensity & gradient strength Gaussians	0	**.85**	.076	±.031
	2	**.78**	.13	±.048
	4	**.76**	.12	±.049
	8	**.85**	.086	±.041
2D Gaussians with covariance	0	**.86**	.059	±.025
	2	**.82**	.071	±.029
	4	**.77**	.087	±.034
	8	**.72**	.11	±.045
Joint Parzen model	0	**.81**	.096	±.037
	2	**.60**	.096	±.038
	4	**.55**	.12	±.051

Table 3: Monotonicity of various objective functions on bladder images. This is the distance to the nearest increasing function from the shape-incorrectness *vs.* objective-function plot, normalized by standard deviation. Thus, the value is zero for a function that is perfectly increasing. The worst it can get is 1.0.

When there is little image blur, different objective functions are equally far from monotonically increasing. But with more blur, the traditional function does not improve; Gaussian PDFs give improved monotonicity; and the Parzen model gives much more improvement.

6. Conclusions and future work

Nonparametric training has proven useful for image segmentation using deformable models. Although the parametric Gaussian model made false assumptions about the distribution of pixel (and differential) values, it still performs quite well. There are undoubtedly unwarrented assumptions even in the "nonparametric" model that performs better, such as the assumption that the estimated distributions of intensity and gradient at the boundary were independent; the choice of a Gaussian kernel; and and human-assisted selection of kernel width; but the model still proved rewarding. Perhaps performance measurement is the only final word on what models are justified.

Though we varied the probability models we used, we did not vary the features they modeled. We are facing another image domain, ultrasound heart images, in which we believe different features will be necessary, and there is no guarantee that multiple features will be independent. We are therefore working on a software framework for multidimensional Parzen-windowed distributions, as well as finding image features that will be effective in distinguishing correct boundaries.

Acknowledgments

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

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.

References

[1] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," in *Proc. IEEE Int'l Conf. on Computer Vision*, (London), pp. 259–268, IEEE Computer Society, 1987.

[2] S. D. Fenster and J. R. Kender, "Sectored snakes: Evaluating learned-energy segmentations," *IEEE Tran. on Pattern Anal-*

ysis and Machine Intelligence, vol. 23, pp. 1028–1034, Sept. 2001.

[3] E. Parzen, "Mathematical considerations in the estimation of spectra," *Technometrics*, vol. 3, pp. 167–190, 1961.

[4] R. P. Grzeszczuk and D. N. Levin, ""Brownian strings": Segmenting images with stochastically deformable contours," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 1100–1114, Oct. 1997.

[5] G. Borgefors, "Hierarchical chamfer matching: A parametric edge matching algorithm," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, vol. 10, pp. 849–865, Nov. 1988.

[6] M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman, "An empirical distribution function for sampling with incomplete information," *Annals of Mathematical Statistics*, vol. 26, pp. 641–647, 1955.