

Segmentation by Posterior Optimization of M-reps: Strategy and Results

Stephen M. Pizer, Robert E. Broadhurst, Joshua Levy, Xioaxiao Liu, Ja-Yeon Jeong,
Joshua Stough, Gregg Tracton, Edward L. Chaney
Medical Image Display & Analysis Group, Univ. of North Carolina

Abstract. For many years we have been developing a variety of methods that together would allow segmentation of 3D objects from medical images in a way reflecting knowledge of both the population of anatomic geometries sought and the population of images consistent with that geometry. To support the probability estimation methods we use to reflect this knowledge, the methods use a medial description, the m-rep, as the object representation and regional intensity quantile functions as the representation of image information in regions relative to the m-rep. Using manually segmented images to which m-reps have been fit and which contain information to allow alignment, our methods use principal geodesic analysis to estimate prior probability density, on the anatomic geometry, and they use principal component analysis to estimate a likelihood density, on the regional intensity quantile functions. They then segment automatically via posterior optimization over principal geodesic coefficients, after initialization via bones or a few contours. Each component of this methodology is briefly reviewed.

Pelvic organs from multi-day populations from individual patients were segmented from CT by training a prior and a likelihood density by the methods indicated. The results are compared to human segmentations. The resulting measurements indicate that in a significant majority of cases, maximizing the log posterior objective function provides segmentations in as good or better agreement with experts than they agree with each other. Similar results are reported for other organs, other image types, and between-patient variation.

1 Introduction

A popular approach for segmenting 3D objects from a medical image in a largely automatic way has been to deform a geometric model into the target image. Such methods optimize an objective function in which a major term measures the match of the model to the target image. Another term can reflect knowledge about the anatomy. While methods using this approach have not achieved the effectiveness to allow broad application to clinical problems, the approach allows the objective function and the feature space over which the deformable model is optimized to reflect two pieces of knowledge humans have when they do manual segmentation: the geometric properties of the anatomic object sought and the patterns of image intensities relative to the object.

A Bayesian point of view allows one to reflect both of these pieces of knowledge. One seeks the most probable object model $\underline{\mathbf{m}}$ given the image \mathbf{I} over objects with the specified geometric properties¹. That is, one optimizes $p(\underline{\mathbf{m}} | \mathbf{I})$ over objects for which $p(\underline{\mathbf{m}})$ is non-negligible. By Bayes' theorem this posterior optimum $\arg \max_{\underline{\mathbf{m}}} p(\underline{\mathbf{m}} | \mathbf{I}) = \arg \max_{\underline{\mathbf{m}}} [\log p(\underline{\mathbf{m}}) + \log p(\mathbf{I} | \underline{\mathbf{m}})]$. In this objective function $f(\underline{\mathbf{m}}) := [\log p(\underline{\mathbf{m}}) + \log p(\mathbf{I} | \underline{\mathbf{m}})]$, the log prior term $\log p(\underline{\mathbf{m}})$ reflects what is known about anatomic geometry and how it can vary. The log likelihood term $\log p(\mathbf{I} | \underline{\mathbf{m}})$ reflects what image intensity patterns are consistent with the anatomy and how they can vary.

Computing $f(\underline{\mathbf{m}})$ for any $\underline{\mathbf{m}}$ involved in the optimization requires training to estimate the parameters of $\log p(\underline{\mathbf{m}})$ and the parameters of $\log p(\mathbf{I} | \underline{\mathbf{m}})$. Making this approach real requires an anatomic geometry representation that can be fit to training data and for which the log prior can be stably trained with the limited number of segmented training image samples that are typically available. As well, the geometric representation must allow intensity data to be considered relative to the represented object. The approach also requires a representation of image intensity data whose log likelihood can be stably trained with a limited number of samples. Finally, it requires optimization and initialization strategies that allow the computation to achieve the posterior optimum. The Bayesian approach helps the computation by limiting the feature space over which the optimization takes place to the relatively few dimensions in which the geometry is adequately typical, i.e., in which $\log p(\underline{\mathbf{m}})$ is adequately large.

¹ We use bold characters to represent vectors or tuples. We use underlines to represent tuples. Thus a bold, underlined character represents a tuple of tuples.

Because they are especially effective for training probability densities, we use the m-rep [Pizer et al., 2003, 2007] as the object representation and discrete regional intensity quantile functions (DRIQFs) [Broadhurst et al., 2006] as the image intensity representation. The m-rep is also strong at allowing the computation of image positions relative to the geometry that is needed to evaluate the consistency of image patterns with the model. Section 3.1 recalls the definition of the m-rep, and Section 3.3 describes how we fit nonfolding m-reps to objects manually segmented from training images. Section 3.2 reviews the definition of the DRIQF.

As motivated in sections 3.1 and 3.2 we use principal component analysis (PCA) on DRIQFs and on linearized representations of the residue of m-reps from their Fréchet mean as the means of estimating probability densities. The PCA on the linearized feature space of m-reps also yields a limited-dimensional space of credible objects within which the optimal object is sought.

As detailed in section 3.5, we initialize the segmentation with the mean model that was computed when training the log prior, globally transformed via a small set of user-provided object-relevant image locations or characteristic relatively rigid image structures, and then we apply conjugate gradient optimization of the objective function over the coefficients of the geometric modes of variation. As described in section 3.4, we optimize the m-rep from large scale level to small scale level, beginning at what we call the object stage and following on to a more local stage, in which the individual m-rep primitives are optimized and for which probability distributions on the primitives are estimated.

Section 4 describes the materials, the methods, and the results of our evaluation experiments. In section 5 we conclude with an interpretation of the results, a discussion of properties of our method, and an indication of developments in progress.

First in Section 2 we survey other deformable-model-based segmentation methods, both to credit methodology that has stimulated ours and to give properties to which those in our method can be compared.

2 Background

Deformable model segmentation methods optimize an objective function over the deformations of a model. For all of them the objective function includes a term that measures how well the deformed model matches the target image. We call such a term a “geometry-to-image match measure” or, for short, an “image match”. One of the earliest methods [Kass et al., 1988] saw the effectiveness of regularizing the objective function by including a term that rewards geometric features that are typical of the objects sought. We call such a term “a geometric typicality measure”.

The earliest methods optimized directly over the geometric primitives of the model, but Cootes et. al (2001) realized the power of having the optimization reflect limited dimensionality obtained from a log prior probability density on the geometric primitives. However, they chose an objective function reflecting only image match and optimized it over a space constrained by the log prior.

Other methods have used an objective function that, like ours, is the sum of a geometric typicality measure and an image match measure but have not used a log probability density for each term [McInerny and Terzopoulos, 1996; Christensen et al., 1994; Joshi et al., 2004]. In this case a “fudge factor” giving a weighting of one of the two terms relative to the other is necessary; its value typically remains as a user choice by trial and error.

Some methods use measures of geometric typicality that integrate local geometric typicality measures over the whole model, and others use image match measures that integrate local image match measures over the whole model. If both properties hold, a local optimization is possible – such methods go under the name “snakes” [Kass et al., 1988; McInerny and Terzopoulos, 1996; Yushkevich et al., 2006]. However, the restriction of geometric typicality to integrated local measures prevents them from richly describing shape, and this has led to methods that too frequently produce segmentations that do not follow what is known about the anatomic possibilities, or that even have improper geometry, i.e., a self-intersecting boundary. An alternative has been to base the geometric typicality on a global representation of the object boundary by coefficients of basis functions [Kelemen et al., 1999] or by log priors computed via PCA on a tuple made of all of the object primitives [Cootes et al., 2001; Tsai et al., 2003; Yang et al, 2003], producing geometric modes of variation that are global. These methods provide no ability to vary the segmentation locally and as a result are subject to the need to optimize over too many coefficients at one time, a process that is not only slow but leads to greater jeopardy of convergence to a local optimum of the objective function.

While many methods have represented the target object directly, typically via its boundary [Cootes et al., 2001; Montagnat and Delingette, 1997], others represent objects via functions on the whole space which are deformed. One major approach computes diffeomorphisms on the whole tuple of voxels [Christensen et al., 1994; Shen and Davatzikos, 2002; Joshi et al., 2004], with objects described by an image whose voxel values are object labels. Another approach embeds the object boundary into a 3D function whose level set gives the boundary [Leventon et al., 2000; Tsai et al., 2003]. The level set approach has the serious advantage of allowing object topology to change as the objective function is optimized. The

segmentation proceeds by solving differential equations on the function. These methods of modifying functions on the whole space require too many parameters to be optimized and thus are subject to slowness and to convergence to local optima of the objective function. The alternative of doing PCA and optimizing over principal vector coefficients has been used, but any deformable model method based on PCA suffers from the instability of PCA on high-dimensional tuples to limited sample sizes [Muller, 2007]. Also, as granted by some of the developers [Tsai et al., 2003], the level set method suffers from the fact that PCA is designed for representations that are vector spaces, a property that level sets do not have. The result is that PCA via level sets sometimes yields geometrically improper segmentations.

Moving on to the measures of image match, two categories of approaches have been used. The most common one integrates voxel-by-voxel matches to a template [Cootes et al., 1993, 2001; Christensen et al., 1994; Joshi et al., 2004], and thus it favors an exact voxel by voxel match of tissue properties as reflected by the image intensities. Alternatively, the template has been set to reflect high contrast without a focus on the particular intensity values, but this has led to bad misses on boundary sections where the contrast is known to be low. The other approach to measuring image match is based on regional tissue mixtures rather than voxel-by-voxel matches. The idea is that in anatomic biology, tissue mixture distributions over regions defined relative to the anatomic model are reasonably constant over the population of images. Moreover, probability distribution estimation over the lower dimensional regional intensity summaries is more stable than over the high dimensional voxel tuples defined over the regions. Such methods have measured image match by regional histograms [Freedman et al., 2005]. The difficulty comes when PCA is applied to the histograms to make the image match probabilistic. Since, for example, the average of two unimodal histograms is typically a bimodal histogram, the linear assumptions underlying PCA are problematic when applied to histograms. We show that a transformation of the histogram, the quantile function, is more appropriate for PCA.

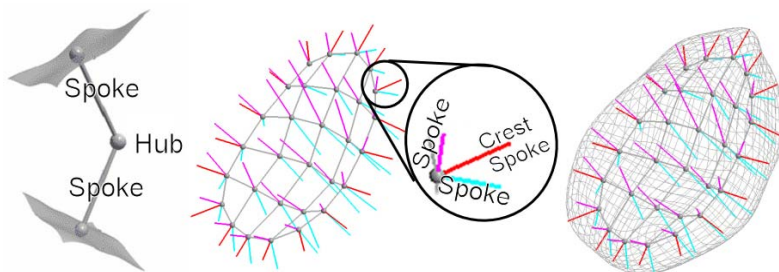
The method we propose uses a log probability density for both the geometric typicality term and the image match term of the objective function, so it has no need for an ad hoc inter-term weighting factor. It operates on many levels of locality (spatial scale), using PCA on residues from larger scale levels aligned at the new scale level to characterize smaller scale information. At each scale level, proceeding large to small, it can optimize over only a few coefficients of principal vectors, yielding speed and limited problems of convergence to local optima. It uses a geometric representation that richly represents shape and can guarantee non-folding shapes. The representation directly identifies object-boundary-relative positions for image intensities used in measuring image match. The image match is based on regional intensity summaries, but both they and the geometric representations of objects are put into a form suitable for the linear analysis of PCA. As described in Section 4, in application to CT images this method is producing segmentations of the prostate, bladder, rectum, kidney, head and neck structures, and subcortical brain structures that by common measures are competitive with human manual segmentation.

3 Method

3.1 The geometric model and probability density on that model

Our ideal model would allow segmentation to be done with multiple levels of locality and allow statistical training of the prior probability density with as few training samples as possible. To most beneficially describe geometry of the object sections that are the focus of smaller levels of locality, the model should also allow one section of an object to be written in geometric relation to its context. The m-rep satisfies all of these needs. In regard to training sample size, we have evidence [Ray et al.] that in certain cases in which the objects in a population are related by local twistings and bendings (rotations) and taperings (magnifications), probability density estimation by a PCA-based method designed to handle such nonlinear transformations on a geometric representation that directly describes these deformations of the interior of the object requires significantly fewer (e.g., half as many) of the expensive-to-obtain segmented images that are needed as training samples.

The m-rep for a 3D object is mathematically defined [Pizer et al., 2003, 2007] as a related group of 2-manifolds with boundary of medial atoms (Fig.1, left), where for all interior points of the manifold with boundary



location \mathbf{p} , the directions, \mathbf{U}^{+1} and \mathbf{U}^{-1} of two spokes, \mathbf{S}^{+1} and \mathbf{S}^{-1} , emanating from the hub to the implied boundary of the object, and the common length, r , of the two spokes. Three scalars are required to represent \mathbf{p} ; two scalars each, $\boldsymbol{\theta}_{+1} = (\theta_{+1}, \phi_{+1})$ and $\boldsymbol{\theta}_{-1} = (\theta_{-1}, \phi_{-1})$,

Fig. 1. A medial atom, a discrete m-rep, and the implied boundary for a bladder.

are required to represent \mathbf{S}^{+1} and \mathbf{S}^{-1} ; and one scalar is required to represent the common length. Thus an 8-tuple is needed to represent an interior medial atom. On the boundary (edge) of each manifold the atom has, in addition to the aforementioned features, a third “crest” spoke bisecting \mathbf{S}^{+1} and \mathbf{S}^{-1} and having an additional length parameter that controls the sharpness of the associated crest on the medially implied object boundary. An edge atom is thus represented by a 9-tuple: the eight scalars of an interior atom plus the additional length.

Each 2-manifold with boundary of medial atoms describes what we call a *figure*. Our m-rep representation $\underline{\mathbf{m}}$ samples this 2-manifold with boundary into a rectilinear grid called a *discrete m-rep* (Fig. 1). Ignoring the crest spoke, each medial atom in the medial grid can be written as a hub translation, two spoke rotations, and a length magnification of any neighbor. In this paper all object representations consist of a single sheet sampled into a single grid, i.e., of a single-figure discrete m-rep. Spatial correspondence between m-reps is given by identifying atoms with corresponding positions in the grid.

A discrete m-rep $\underline{\mathbf{m}}$ with e edge atoms and i interior atoms can be understood as a point in a $9e+8i$ -dimensional feature space. The feature space must be understood as a curved surface due to the fact that the spokes are described by an angle of rotation [Fletcher et al., 2004]. Relating changes in the various atom components via their instantaneous effects at the implied object boundary, i.e., at the spoke end, yields the following expression for the distance between a medial atom \mathbf{m} and its value after deformation: $|\Delta\mathbf{m}| = [|\Delta\mathbf{p}|^2 + |\bar{r} \Delta\theta_{+1}|^2 + |\bar{r} \Delta\theta_{-1}|^2 + |\bar{r} \Delta\log r|^2]^{1/2}$, where $|\Delta\theta_{\pm 1}|$ is the length of the geodesic path on the unit sphere between $\theta_{\pm 1}$ and its value in the deformed object, and where \bar{r} is the mean of the spoke length in that atom in a population (see below). Also, if a discrete m-rep $\underline{\mathbf{m}}$ consists of medial atoms \mathbf{m}_i , $|\Delta\underline{\mathbf{m}}| = [\sum_i |\Delta\mathbf{m}_i|^2]^{1/2}$.

Estimating a probability density $p(\underline{\mathbf{m}})$ from a training set of N samples $\underline{\mathbf{m}}^k$ of a discrete m-rep requires first computing the mean $\underline{\mu}$ of the n samples. $\underline{\mu}$ must be computed with respect to the measures of distance above. That is, since the ordinary formula for mean does not apply in curved spaces, we use the Fréchet mean, namely the point in the feature space of m-reps that has minimum sum of squared geodesic distances to the sample m-reps. This yields an m-rep each of whose atoms μ_i is computed as follows:

- 1) Its hub is the ordinary mean $\frac{1}{N} \sum_{k=1}^N \mathbf{p}_i^k$ of the hubs \mathbf{p}_i^k of the corresponding atoms in the training set;
- 2) Its atom spoke length \bar{r}_i is $\exp[\text{the ordinary mean of the } \log r_i^k \text{ values}]$, i.e., $\exp[\frac{1}{N} \sum_{k=1}^N \log r_i^k]$;
- 3) its spoke directions $\mu_{U^{\pm 1}_i}$ are the Fréchet mean on the unit sphere of the sample spoke directions of the corresponding spoke.

For crest atoms the crest spoke length in μ_i is also computed by the same means as applied to the paired spokes in item 2.

At the m-rep feature space point $\underline{\mu}$ corresponding to the Fréchet mean on this curved feature space, there is a closest fitting linear space of the same dimension tangent at that point. Projections of points on the curved space onto this tangent space produce a set of points on which PCA can be applied. The PCA yields a set of orthogonal principal vectors \mathbf{v}_i on the tangent plane and a corresponding set of principal variances, σ_i^2 . Taken together this yields a Gaussian $p(\underline{\mathbf{t}})$ with mean $\underline{\mu}$ on the tangent space such that for $\underline{\mathbf{t}}$ in the subspace of the tangent space spanned by the \mathbf{v}_i , i.e., for $\underline{\mathbf{t}} = \underline{\mu} + \sum_i a_i \mathbf{v}_i$, $-2 \log p(\underline{\mathbf{t}}) = \sum_i a_i^2 / \sigma_i^2 + \text{a constant}$. Since any vector in the linear space with its tail at the mean can be projected back down onto the feature space of m-reps and the Euclidean length of the vector is equal to the geodesic length of the projected vector, this corresponds to generating a probability distribution on m-reps. Specifically, let $\underline{\mathbf{t}}$ be a point on the tangent space, with corresponding m-rep $\underline{\mathbf{m}}$ on the curved feature space, such that $\Delta\underline{\mathbf{t}} = \underline{\mathbf{t}} - \underline{\mu}$ can be written as a linear combination, with coefficients a_i , of the principal vectors chosen through PCA on these tangent-plane-projected residues; in the shape space of these principal vectors $\underline{\mathbf{t}}$ is described by the vector \mathbf{a} . Then $-2 \log p(\underline{\mathbf{m}}) = \sum_i a_i^2 / \sigma_i^2 + \text{a constant}$. This analysis in terms of a Fréchet mean and PCA on the tangent plane at the Fréchet mean is called Principal Geodesic Analysis (PGA).

To obtain the probability distribution tightness that yields stable estimation of the distribution using small sets of training samples, the training models need to be pre-aligned.

3.2 The intensity model and probability density on that model

In this paragraph we argue that when using probabilistic measures to form an image match, having those probability distributions be on regional intensity summaries is superior to having them combine voxel-by-voxel intensity values. Intensities in an anatomic region of some scale vary according to four factors: 1) the imaging device settings, 2) the random noise in imaging, 3) the textures of the physical variables of each tissue type in the region, and 4) the mixture of tissue types in the region. With a geometric representation of regions associated with a deformable geometric model, we can assume that the mixture of tissue types in the region has small variability. Thus combining the four factors by considering distributions of intensities within the anatomic region can encompass small variability and thus more stable trainability, as compared to

considering measurements based on subregions comparable to or smaller than the scales of the texture, e.g., specific voxels' intensities. Therefore, spatial correspondence may best not be refined below the texture scale to give voxel-to-voxel correspondences. On the other hand, having the regions be so large that they do not capture locality of the tissue mixtures will yield an image match that is inadequately spatially descriptive. Finally, as long as the tissue mixtures in a region are relatively stable, rather than using too small regions it may be better to err in the direction of regions that are somewhat too large, providing more voxels to generate statistics.

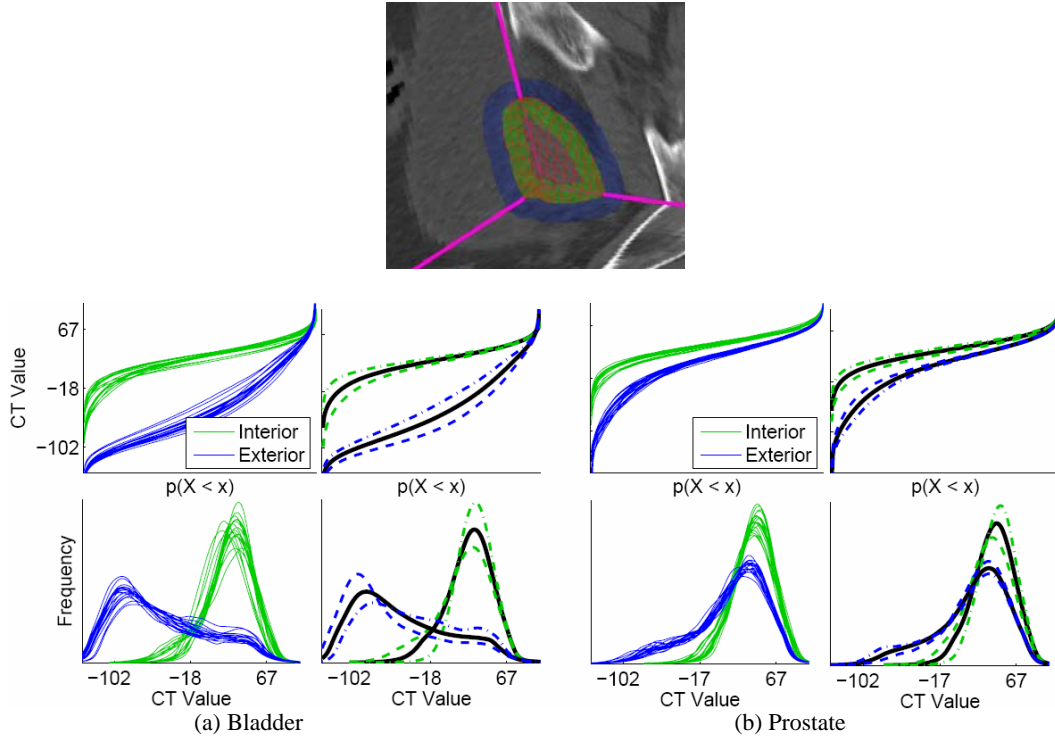


Fig. 2. Top: prostate regions for RIQFs: the grid shows the prostate surface superimposed on a tri-orthogonal display of the image data. The exterior region on those image slices is shown in brown and the interior region in blue. Middle: Bladder and prostate training RIQFs for interior and exterior regions. For each, the left panel shows the training samples, and the right panel shows the learned mean and ± 2 standard deviations along the first principal direction. Bottom: Histograms corresponding to the respective RIQFs.

One can stably summarize the collection of voxel intensities in a region by their probability density function (intensity histogram), or equivalently their intensity quantile function (IQF), which gives for each quantile between 0 and 100% the intensity at that quantile. The IQF is the inverse of the cumulative distribution function (CDF). Across many instances of the same region, e.g., across patients or within a patient across days, these IQFs have their own variability. As explained below and illustrated in Fig. 2, if the dominant variability of the IQFs is from the factors 1-3 listed above, regional IQFs (RIQFs) lend themselves to PCA after discretization. Thus our primary representation of a region's intensity pattern is a discrete sampling of the quantile function, which we term a discrete regional intensity quantile function (DRIQF) [Broadhurst et al., 2006]. A DRIQF is easily computed by sorting the intensities and averaging adjacent values, forming a set of quantile averages. The DRIQF representation is much less sensitive to over-binning than histogram based approaches since instead of bin counts DRIQFs store bin positions in the stable list of sorted intensities [Broadhurst et al., 2006].

As is discussed in detail in [Broadhurst et al., 2006], there are two main reasons to consider the space of DRIQFs as Euclidean, and thus meeting the assumptions for PCA. First, Euclidean distance corresponds to a metric known as the Earth Mover's, or Mallows, distance [Levina and Bickel, 2001], which is an intuitive measure of the difference between distributions. Second, any affine transformation of the variable described by the DRIQF (here image intensity) yields a linear change in the DRIQF, allowing DRIQF populations to be analyzed by linear statistical methods such as PCA; such affine transformations include changes in the distribution's mean and standard deviation. A theoretical drawback, however, is that DRIQFs do not form a vector space; only a convex subspace is valid. A more practical limitation is that though many forms of distribution variation are linear, other important forms of variation are nonlinear, including changing the mixture amount in a distribution composed of a mixture of two widely separated peaks. Nevertheless, as compared to probability distributions on the ordered tuple of voxel values in the regions, the probability distribution on a DRIQF is more informative both because

it is less variable, as discussed at the beginning of this section, and because it is more stably estimated, due both to its far lower tuple length than the ordered tuple of voxel values and to the far lower number of dominant principal eigenmodes, as just discussed.

Due to both the strengths and the weaknesses of DRIQFs, for our final regional intensity representation within CTs we threshold out extremely low and high CT values, corresponding to gas and bone tissue, respectively. For each region k , this produces three separate intensity distributions: A_{gas}^k and A_{bone}^k , representing the voxels with thresholded low and high intensities, respectively, and \mathbf{Q}^k , representing the other voxels in the region. We then independently model the amount of each of these intensities in a region. That is, for region k , $p(\mathbf{I}^k) = p(A_{\text{gas}}^k) p(\mathbf{Q}^k) p(A_{\text{bone}}^k)$. We represent \mathbf{Q}^k by its DRIQF and estimate $p(\mathbf{Q}^k)$ using PCA, yielding dominant principal vectors, a principal variance for each vector, and the residual variance over and above the selected dominant modes. We estimate $p(A_{\text{gas}}^k)$ and $p(A_{\text{bone}}^k)$ by assuming A_{gas}^k and A_{bone}^k follow a Gaussian distribution with an enforced minimum standard deviation.

We currently assume that the voxel intensity patterns within each region are independent of those in the other regions, yielding $p(\mathbf{I} | \mathbf{z}) = \prod_{k=1}^K p(A_{\text{gas}}^k) p(\mathbf{Q}^k) p(A_{\text{bone}}^k)$ and thus, the image match, $-2 \log p(\mathbf{I} | \mathbf{z}) + \text{a constant}$, =

$$\sum_{k=1}^K \left[\left(\frac{A_{\text{gas}}^k - \mu_{\text{gas}}^k}{\sigma_{\text{gas}}^k} \right)^2 + \sum_i b_i^{k^2} + \left(\frac{d_{\text{proj}}^k}{\sigma_{\text{proj}}^k} \right)^2 + \left(\frac{A_{\text{bone}}^k - \mu_{\text{bone}}^k}{\sigma_{\text{bone}}^k} \right)^2 \right], \text{ where } b_i^k \text{ are } \mathbf{Q}^k \text{'s coefficients of the } k^{\text{th}} \text{ region's chosen principal vectors (the}$$

region depends on the vector \mathbf{a} specifying the object \mathbf{z}), $\tau_i^{k^2}$ are the corresponding principal variances, d_{proj}^k is the projection error (or residue) of \mathbf{Q}^k onto the space defined by the principal vectors, and σ_{proj}^k is the expected projection error.

We have tried two different divisions of the region near the object boundary into regions. Both separate the region interior to the medially implied object boundary from that exterior to the boundary, and both use an interior region that is assumed to be homogeneous enough in tissue mixture that it needs no subdivision. One division uses a global exterior region, whereas the other provides locality in the exterior, and thus accommodation of exterior inhomogeneity due to different anatomical structures surrounding the target structure. The exterior is thus divided into as many subregions as there are spokes in the model (typically, many tens) according the placement of the medial atom spokes. Using multiple exterior regions produced better results on our bladder and prostate segmentation tasks, so the results reported later for these organs used these subregions. In each region the contribution of each voxel's intensity into a DRIQF is Gaussian weighted by its distance to the object boundary, which provides a smoother objective function and allows narrow regions to be defined with larger capture ranges. We use a Gaussian with a standard deviation of 3mm, 100 samples in a DRIQF (an insensitive parameter), and low and high threshold values of -224 and 176 Hounsfield units to threshold out gas and bone.

We have found it effective to modify the objective function by making each term have the same variance. We do this by transforming each term, which follows a χ^2 distribution with some number of degrees of freedom, M , into a mean 0, variance 1 random variable by subtracting its mean, which is equal to M , and dividing the result by the formula for its sampling standard deviation, which is $(2M)^{1/2}$.

3.3 Training the log probability densities

Training both the log prior and the log likelihood begins with a collection of training images in which manual segmentations of the objects of interest yield binary images. M-reps are fit to each of these binary training images [Merck et al., 2006] using a variant of our segmentation program in which the geometry-to-image match depends on distances between the m-rep's implied boundary and the binary object's boundary, and in which the m-rep's geometric typicality depends on the geometric propriety (non-foldedness and even smoothness) of the implied object using the medial mathematics of [Damon, 2007], the regularity of the m-rep's atoms, and possibly the distance of those atoms from those of a reference m-rep, after alignment [Han et al., 2007]. Our fitting method produces objects whose appearance is qualitatively good and that match the binary data with submillimeter accuracy even when the voxels are $1 \times 1 \times 3$ mm (see Fig. 4), but this accuracy together with robustness of fitting only held after the geometric propriety penalty was applied.

Training the log prior. We apply principal geodesic analysis (PGA) [Fletcher et al., 2004] on m-reps to produce the Fréchet mean and the set of unit principal direction vectors \mathbf{v}_i and principal variances σ_i^2 from which a dominant subset is chosen (4-8 for the organs studied in this paper). As described in section 3.1, target object's with \mathbf{a} as the coefficients of the principal vectors has a Mahalanobis distance log prior penalty of $\sum_i a_i^2 / \sigma_i^2$.

Training the log likelihood. The fitted m-reps define object-relative image regions. Each region's voxels are identified by following normals from the m-rep implied object boundary points. The voxels are then used to measure the amount of gas and bone tissue intensities and a discrete quantile function. PCA is then applied to the discrete quantile functions with two dominant modes chosen for the interior region and three for each exterior region. Two and three modes were chosen so as to account for most of the variance, at least 95%, while leaving roughly the same amount to absolute residual variance, which reduces interior/exterior bias in the segmentation. We only include voxels in the respective regions that have the correct label in the manual segmentation (1 for inside, 0 for outside). This defines an ideal optimum of the likelihood but it skews the expected variation from this optimum since during target time the expected m-rep defined segmentations will contain errors. To correct for this, we rescale the principal variances so that the average Mahalanobis distance of the training regions without the manual segmentation correction is equal to its expected value, the number of principal modes plus one for the expected projection error.

3.4 The objective function at multiple scale levels

The objective function described above applies at the level of the object, where the segmentation optimizes over the coefficients of the selected principal geodesic modes of the tuple of atoms representing the object. An objective function of the same form on atom residues from the object stage results applies to refine the segmentation, atom by atom. However, its log prior must be on locally aligned atom residues. We align relative to the orientations of the medial normal of the atom and its immediate neighbors.

3.5 Initialization

All deformable model methods on medical images are sensitive to initialization. Four different initialization methods of the mean m-rep produced in training have been used in the studies reported below.

- 1) User-provided topmost and bottommost and central axial contours, with optimization of object eigenmodes of variation, i.e. warps. The Mahalanobis distance providing the geometric atypicality penalty is then computed from this initialized object.
- 2) Within-patient similarity transform based on intensity matches of the high intensities that identify the bones (in the male pelvis).
- 3) Within-patient or between-patient similarity transform based on landmarks. This is fine for alignment of training cases but really is unsuitable for clinical initialization, as for most target organs users cannot be expected easily to identify relevant landmarks in 3D.
- 4) Between-patient similarity transform based on the global containing structure (skull and whole brain) of the target structures. This method is suitable for the brain, with its relatively stable position of substructures relative to the global structure.

4 Experiments and Results

The objectives of the research reported here were two-fold: to determine the effectiveness of the -log posterior objective function that we use and to measure the quality of the combination of initialization, object stage optimization, and atom stage optimization in producing segmentations. We begin with segmentation of the prostate and bladder from CT on a given patient varying his geometry from day to day. The experiment on these materials are reported in some detail. We then briefly summarize results on a variety of organs from a variety of images with between-patient variation and on a largely tubular object, the rectum, from the same CT scans as in the main experiment.

4.1 Prostate and bladder with within-patient variation

Multi-day CT scans of at least 14 images of the male pelvis region acquired for image-guided radiation therapy for 5 patients were selected for study; there were a total of 80 cases. Four patients were from UNC, and one was from William Beaumont Hospital. The prostate and bladder, and a fixed length of rectum were manually contoured slice by slice in all

scans. Geometric object stage probability distribution estimation was performed on the fitted m-reps after alignment. The alignment used for statistical training was via bone-based similarity transforms for both the prostate and the bladder.

Leave-one-out experiments were carried out for the prostate and the bladder separately, whereby in each case within the experiment training of the log prior and log likelihood was done on all but one of the images (days) within the respective patient's set and segmentation was carried out on the remaining image. The target images were 3D CT images with $1 \times 1 \times 3$ mm voxels. For the prostate we initialized by a similarity transform that optimized the match between the pelvic bones to the first treatment day. For the bladder we tested the optimization by initializing the mean target object in the object stage onto the target image first via the just described bone-based similarity transform, followed by a shape space warp determined by the three contours described in section 3.5, item 1. After the initialization, we applied the object stage and then the atom stage.

The segmentation was compared to the corresponding human segmentation using two measures: intersection/average volume overlap (DICE) and average closest point distance. The pelvic organ results combine segmentations over 5 patients and a total of 80 images (on the average 16 days of each patient), with that patient's not-left-out days providing the training for that patient's left out day. The human segmentations compared to were from the same human whose segmentations formed the training in the left-out cases.

The primary result (Figs. 3, 4) is that by the measures of average distance to the closest boundary point and volume overlap to the human segmentations, our method's results are distinctly superior to any others reported for the segmentation of these organs from CT (e.g., see [Costa et al., 2007 and its reference list; Jeong and Radke, 2006]). Moreover, not only are the average distance measurements subvoxel, but for 77 of the 80 cases the average distance of the prostate segmentation to the trainer was less than or equal to the median case average distance of the segmentation of another expert to that of the trainer, namely 1.9 mm [Foskey et al., 2005]. No bladder had an average distance greater than 1.6 mm.

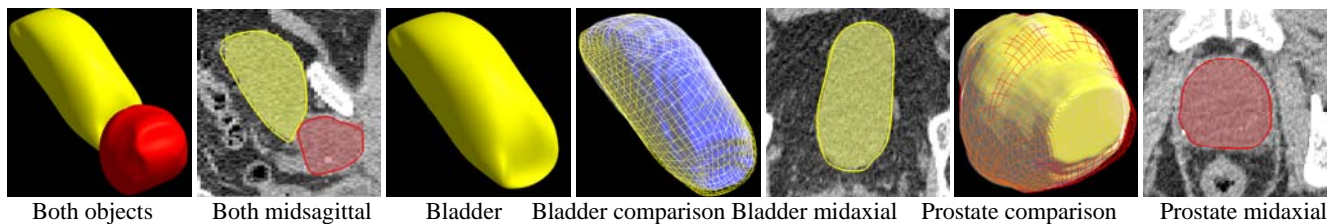
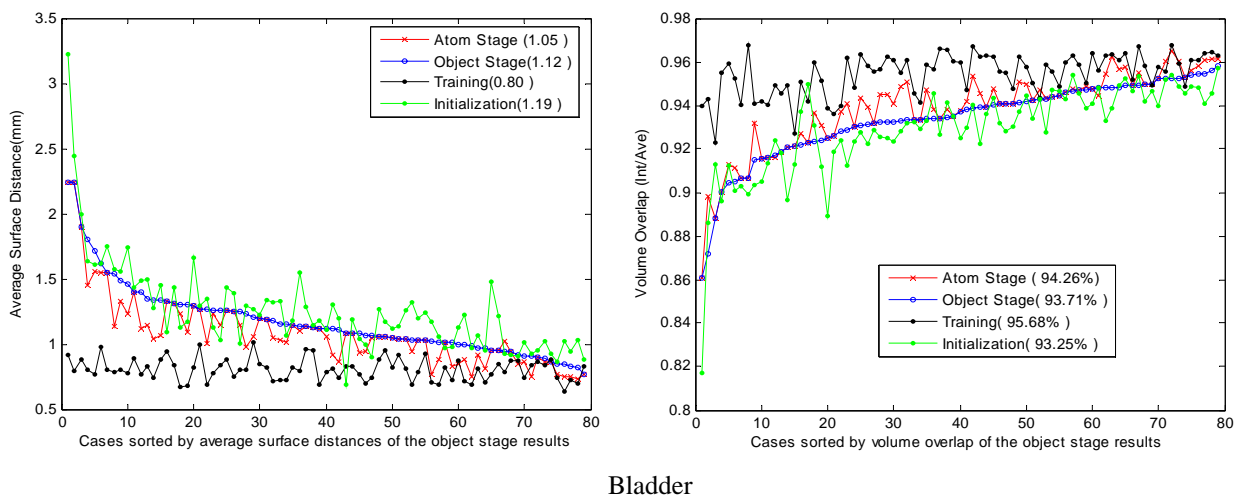


Fig. 3. Sample segmented bladder and prostate in a typical case. From left to right: surface rendering of segmented bladder and prostate, midsagittal slice with human segmentations colored and computer segmentations as curves, surface rendering of segmented bladder, computer segmentation of bladder in wire mesh vs. surface rendered human segmentation in blue, surface rendering of segmented prostate, computer segmentation of prostate in wire mesh vs. surface rendered human segmentation in blue, midaxial slice of bladder with human segmentation colored and computer segmentation as curve, midaxial slice of prostate with human segmentation colored and computer segmentation as curve.



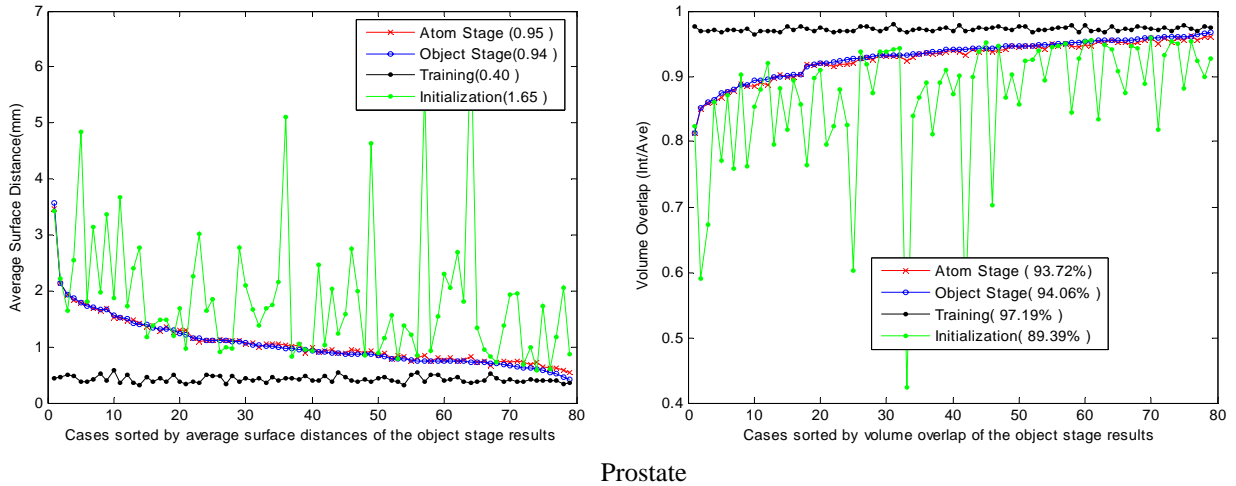


Fig. 4. Average distance and volume overlap for segmentation of the bladder and the prostate in 80 cases, where the training was on approximately 15 cases from the same patient. For each graph, the results of the object stage are sorted in increasing order of quality and the results for the training segmentation for that case, for initialization for that case, and for the results of the atom stage for that case are shown opposite the object stage result with the same case number. The small box shows the median value for the corresponding measure.

For the bladder we found that, as expected, in regions of reasonable contrast where the object stage left the segmentation a voxel or two off, the atom stage brought the segmentation to subvoxel accuracy. For the prostate, again as expected with the low contrasts available in CT, the atom stage did not yield improvement but did not hurt.

Testing the effectiveness of the $-\log$ posterior objective function used at the object stage or the objective function, including the log likelihood, used at the atom stage requires looking beyond the local minimum achieved by conjugate gradient optimization. All geometry-to-image match penalty functions, including ours, appear to have many local minima, and geometric typicality penalty functions, including ours, are convex. Hence objective functions for deformable model segmentation are typically bumpy. This results in optimizers having some difficulty in finding global minima. To obtain the lowest minimum we could find, we could drop the requirement of a clinically possible initialization. Thus, for each target case we also initialized the optimization at the m-rep that was fit to the binary image that was the manually produced “right answer”. More precisely, for the pelvic organs, where only the object stage was applied, the alternative initialization was the projection of the “right-answer” m-rep on the principal geodesic space of the training. In most cases, the original result was not significantly different according to distance or volume overlap measures from the ideally initialized result, even if a different objective function minimum was reached.

The effect of the original initialization was also studied via the average boundary distances and volume overlaps of the initialized m-reps relative to the human segmentations and via the relation between the distance or overlap pattern on the object to those patterns after the segmentation. Fig. 4 shows that the contour-based bladder initializations are already, in median, at 1.2 mm average distance from the human segmentations. This is already distinctly better than many final segmentations reported in the literature. The fact that the initialization based on optimization over the shape space via the weak data of 3 contours is so close to the desired result shows how important the calculation of a good shape space is. In the case of the prostate the initialization did not include a shape space warp, and this resulted in a median average distance of initialization of only 1.65 mm.

The segmentation steps improve the average distance measure over the initialization by only about $\frac{1}{4}$ mm in the majority of the bladder cases and by about $\frac{3}{4}$ mm from the poorer initialization in the case of the prostate. In some cases the segmentation improved the average distance by as much as 4 mm. Segmentation improved the relatively insensitive measure of volume overlap over initialization in the median case by just over 1% of the volume for the bladder and by around 4% for the more poorly initialized prostate in the majority of the cases. Roughly, the better the initialization, the better the final result.

As shown in Fig. 6, even for the m-rep fits to the binary training segmentations the Hausdorff distance error over the boundary in the median case is 3.5 mm for the bladder, which is an elongated structure, but it is only 2.4 mm for the smaller, more blob-shaped prostate. The Hausdorff distance error drops by only $\frac{1}{4}$ mm for the bladder but by 1 mm for the prostate between initialization and segmentation.

4.2 Segmentation of other organs from various images

Kidney with between-patient variation. 39 slice-by-slice-scanned CTs of different patients with completely imaged kidneys without pharmaceutical contrast were acquired. The left kidneys were carefully segmented slice by slice using interactive contouring tools. 6 landmarks were identified for each kidney: 2 at the north and south poles, and 4 on a belt around the midsection of the kidney. When m-rep models were fit to the manual contours for geometric training as described in 2.3, the ends of pre-identified spokes were forced via a penalty in the objective function to correspond to the landmarks within a preset tolerance to achieve anatomic correspondence and alignment across the training m-reps and to provide a means for initialization of the mean m-rep in target images. A leave-one-out experiment similar to the one described for the pelvic organs was carried out in 2006, using a version of our method that was available two years ago. This experiment followed earlier experiments on different kidney CTs [Rao et al., 2005] that showed kidney segmentations that, relative to segmentations by the trainer, were competitive with another expert's segmentations relative to the trainer.

In the 2006 experiment, we compared segmentations using two likelihood functions (image match measures) while fixing the geometry probability distributions. One of the likelihoods compared was based on voxel-by-voxel profiles normal to the boundary [Stough et al., 2004], and the other was based on global RIQFs (one interior region and one exterior region). The results showed that segmentations using the RIQFs [Broadhurst, Stough, et al., 2006] improved the average distance to the manual segmentations by 0.4 mm on the average, to an average distance of 1.8 mm. Today's methods based on local interior and exterior RIQFs and improved m-rep fits in training give better results still, at least for segmenting prostates and bladders.

Head and neck with between-patient variation. In a leave-one-out study using 8 between-patient CT studies (all we have with careful human segmentations so far) the sterno-cleido-mastoid and masseter muscles and the parotid gland were segmented using landmark based alignment for both statistical training and for initialization. For those target cases where the m-rep fitting the manual segmentation was an in-lier with respect to the other 7 cases used for training, the segmentation at the object stage agreed with the manual segmentation by an amount consonant with the results reported above for pelvic organs.

Subcortical brain structures with between-patient variation. In a study using a between-patient population of 120 MRI images in a study of neurodevelopment in autism, the left caudate nucleus, globus pallidus, hippocampus, and putamen were segmented at the object stage followed by the atom stage. Fifty of the images were used for training. The initialization was simply at the mean model position in the MRIs, which had already been rigidly normalized on the basis of the whole brain and skull. Before training and segmentation each MRI, with 0.8 mm isotropic voxels, underwent voxel classification by statistical pattern recognition [Prastawa et al., 2003] into CSF, grey matter, and white matter classes and then was intensity-normalized to bring the means and standard deviations of the histograms from each of these three classes to be as standardized as possible for each case. The results were again subvoxel average distances in segmentation against careful, highly repeatable human segmentations [Hazlett et al., 2006]. Fig. 5 shows an example segmentation the caudate.

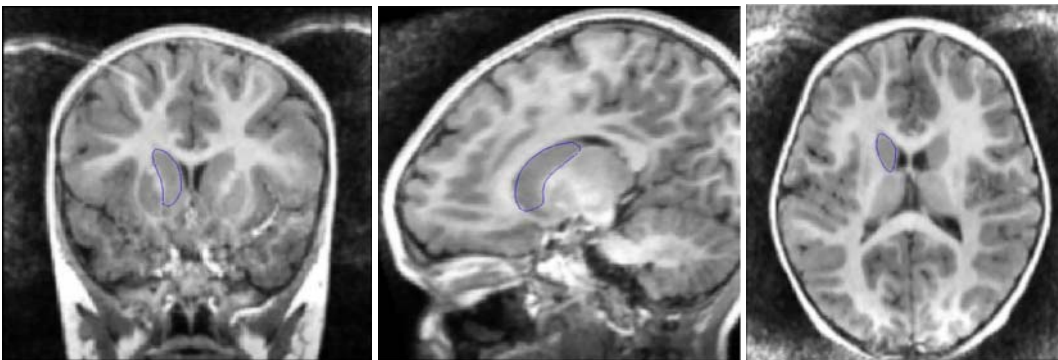


Fig. 5. A typical caudate segmentation from MRI, with the curve showing a slice of the automatic segmentation. Left: axial view. Middle: sagittal view. Right: coronal view.

Rectum with within-patient variation. Because nearly tubular objects (a tube has circular cross sections) are unstable in the orientation of their medial slab, our early experiments indicate it is advantageous to model the rectum as a tube with refinements taking care of the deviations from tubularity. We [Saboo et al., 2007] have developed m-rep tubular models, consisting of chains of medial atoms formed from a cone of spokes, methods for estimating their probability distributions and

probability distributions of RIQFs for tube-relative regions. While our experiments of segmentations of the rectum are still in progress, the results to date of the object stage are giving promising results.

5 Other Results, Interpretation, and Conclusions

5.1 What components of our method lead to its success?

Our method is based on estimating probability distributions on anatomic geometry and on anatomy-relative intensity patterns, and as such is quite generally applicable across anatomic targets and image types. We choose geometric and intensity pattern representations so that their probability distributions will be stably estimated with relatively few training cases, leading to the choice of the m-rep and the RIQF as the respective representations. In addition, we choose the m-rep because of its ability to represent the relations between geometric entities at different locations in terms not only of relative position but also relative orientation and thus to firmly support multi-scale optimization, leading to greater likelihood of convergence to the global optimum of the posterior probability, which is its objective function.

It is hard to know in which part each of these properties lead to its producing very good segmentations. However, eight pieces of evidence are relevant.

- 1) As reported in section 4.2 with the kidney study, as a part of our overall method the RIQF provides superior performance to voxel-to-voxel scale based measures of image match. This provides both more stable estimation of the likelihood function, due to lower dimensionality, and an association with tissue properties at what appears to be a more appropriate scale.
- 2) In an early version of our method we compared two choices for geometric typicality, with both using the statistically trained voxel tuple method as the geometry-to-image match referred to in item 1. Both methods used the geodesic mean as the model and optimized over the coefficients of principal geodesics. One of the methods used the PGA-based Mahalanobis distance as the geometric typicality penalty, and the other used a geodesic distance between the candidate and the mean, i.e., a geometric, nonprobabilistic method. The Mahalanobis distance produced superior segmentations. Together with the stronger argument of the high quality segmentations we get using a log posterior as compared to other methods reported in the literature, this argues for probabilistic geometric typicality functions over deterministic ones.
- 3) In an earlier version of our method we compared two choices for image match, with both using optimization over log prior PCA components and a log prior geometric typicality [Broadhurst et al., 2005a]. The comparison was on bladder and prostate segmentation in a single patient. In one alternative the image match was the Earth Mover's distance to the mean, and in the other it was the Mahalanobis distance to the mean. Better segmentations were produced using the Mahalanobis distance.
- 4) Our method improved its performance each time we improved the quality of m-rep fits to binary training images and thus the quality of both probability distributions on which the method is based.
- 5) In our experience our method produces more frequent convergence to a satisfactory relative optimum than methods based on diffeomorphism at the voxel scale, and it does so in much improved computing times of ten minutes or so. This is based on our working at the object scale and then the object part scale and not only at the voxel scale.
- 6) We have studied the effect of accuracy of credible variation of the initialization on the final segmentation's accuracy. Roughly, in earlier versions of our method, what we found was that the particular optimum achieved varied with the initialization but that frequently even the local optimum at which convergence stopped produces a segmentation not far (in the image space) from the ideal segmentation.
- 7) Initializations in the space of credible models and thus involving warps of the mean model have improved the segmentations of an object with strong geometric variability, the bladder.
- 8) Aligning the training m-reps via two manually placed prostate landmarks, at the urethral entrance to and exit from the prostate, produced geometric probabilities that yielded segmentations negligibly (~ 0.05 mm) better in the median case than those produced with alignments based on the pelvic bones.

Another important advance in our method followed attention to fitted models having proper geometry in training [Han et al., 2007]. Before we penalized geometric legality in training, segmentations would sometimes be created and concomitantly produce bad fits to the image data due to the incorrect boundary normals on which the fits were based.

Handling tissues with identifiable, extreme intensities specially (gas, bone) has improved the performance of our method.

5.2 What behaviors of our method need improvement?

Though it would seem from Fig. 6 that maximum errors of our method's segmentations from the human segmentations are typically too great to permit avoidance of editing, in fact the large majority of segmentations (e.g., 77 of 80 for the prostate) are clinically usable. Fig. 7 illustrates our findings that explain the situation. First, as the worst case bladder case shown in the left pair of images in Fig. 7 illustrates, the large difference from the human's segmentation is in a region where the object boundary location is equivocal; the automatic segmentation is as clinically usable as the human segmentation. Second, as is the case for both the worst case bladder and the worst case prostate shown in the right pair of images in Fig. 7, the source of the error is poor initialization. In the case of the bladder, the poor initialization and the consequent errors in segmentation was due the fact that it had the largest volume of all of the days for that patient and thus was an outlier with respect to the trained probability distribution of geometry used in the segmentation. We are impressed with what a large fraction of usable segmentations were produced with only 15 training cases on the average, but to train for clinical usability, cases encompassing a wider range of situations will be needed. In section 5.3, we discuss the plans for training for clinical use.

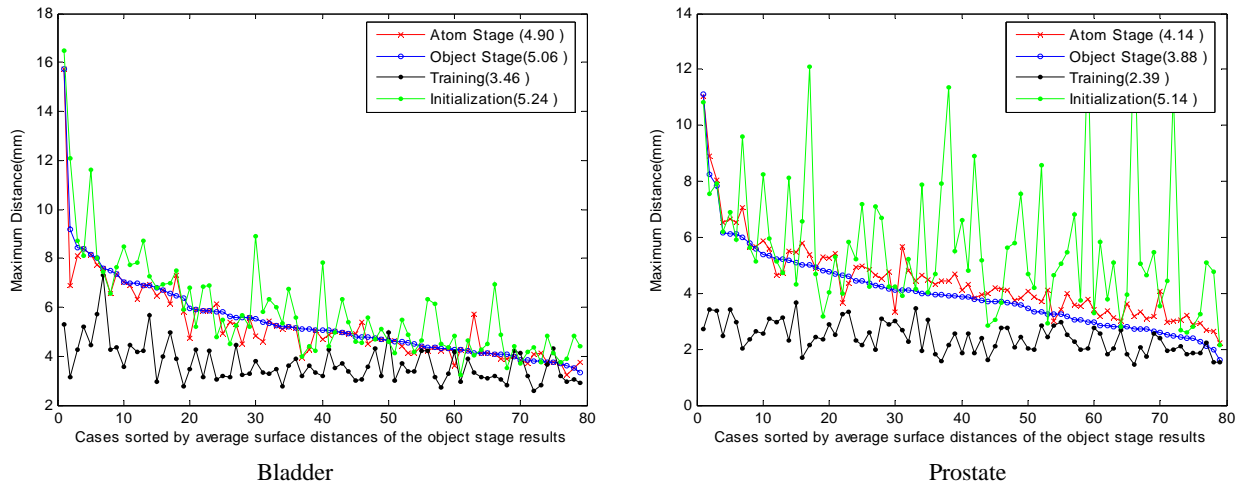


Fig. 6. Maximum (Hausdorff) distance for segmentation of the bladder and the prostate relative to the human segmentation in 80 cases. Left: bladder; right: prostate. The small box shows the median value for the corresponding measure.

The worst case prostate in terms of maximum distance was initialized in an intensity environment that is rather consistent with common intensity patterns seen in training. In that case the bone position was a poor predictor of the prostate. In the worst cases for the bladder and prostate shown in Fig. 7, both of the errors had the property of having a significant overlap between the prostate segmentation and the bladder segmentation, done independently. Jeong et al. [2006] in our group are developing a method of prediction of one organ via statistics of its interrelation of another neighboring organ, more precisely as the conditional mean of the predicted organ's m-rep given nearby medial atoms in the neighboring organ's m-rep. When one of the objects is reasonably well segmented, as with the bladder, even from its 3-contour initialization, the prediction of the neighboring object, here the prostate, has turned out to be quite good in initial studies. With this improved initialization, many of the worst cases of Hausdorff distance would be avoided.

In the remaining cases where there is a patch with a large error, we believe the computer must signal the non-credibility of that patch as often as possible. Joshua Levy is developing methods for testing patches' RIQFs being an outlier within its probability density [Levy et al., 2007] to identify the patch as needing editing.

Sometimes the reason for the maximum error is that convergence is to a local optimum of the objective function. This is a standard problem with deformable models methods. An aspect of this is sensitivity to initialization. Local optima abound in most measures of image match, including ours, and methods for climbing out of a local optimum of our objective function are worth investigating. We have begun such investigations.

For the objects we have investigated so far, the geometric probability distributions have shown themselves to be unimodal in our examinations. However, there is a possibility of multi-modal shape probability distributions, even of normal organs; for example, we have encountered organs (e.g., the parotid on one side of the face) that come in single-object and two-object forms. Consideration of how to handle multimodal probability distributions will be worthwhile.

Of course, medicine deals not just with normal anatomy or pathologies with subtle geometric changes from normal, with which our method is designed to deal. Handling pathology that leads to gross changes in shape must be faced, and this will be a challenge due to the wide variability of shape changes that pathology can sometimes engender. In the case of tumors, we

need to make use of the physical models of tissue deformation that are available. We can make use of the FEM multiscale meshing and computation methods based on m-reps reported in [Crouch et al., 2007].

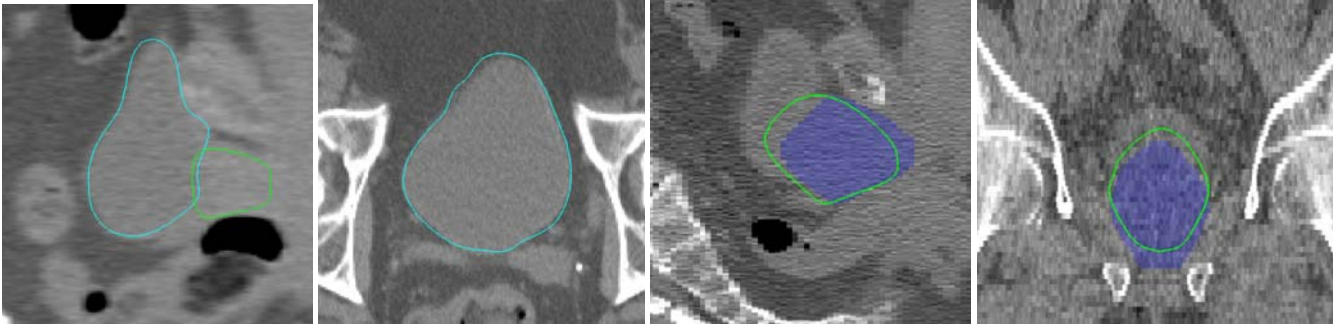


Fig. 7. Case and slices with worst maximum error for segmentation of the bladder and prostate. Left: midsagittal view of both organs from worst-case bladder. Middle, left: midaxial view of worst-case bladder. Right pair: Midsagittal and axial views of worst-case prostate, with human segmentation in color wash.

Other deformable models methods have benefited from improved correspondence based on regularity and entropy measures [Davies et al., 2007; Cates et al., 2006]. M-reps, by combining orientation, size, and position information and with our using closeness of atoms to a reference (tentative mean) model in training as a reward in the objective function used in fitting, we may give better correspondences in their fits than some other representations. Nevertheless, we anticipate that improving correspondences in training will be helpful.

There are a number of extensions of the method on which we are already embarked. They are listed and discussed in section 5.4.

5.3 How can we make our method clinically usable?

In the adaptive radiotherapy problem that stimulated our within-patient segmentation of male pelvic organs, on an early day of treatment there are not enough previous days to provide adequately trained probability distributions of m-reps and of RIQFs. In [Pizer et al., 2006] we have described a variant of our method in which the means used in the statistics are the means of the previous days for the target patient but the variations come from pooling the differences from their all-day means of other patients, after appropriate between-patient alignments. In a slightly earlier version of our method we showed that, as compared to the method reported in section 5.1 that entirely uses within-patient learning of the probability distributions, this approach gives segmentations with equal robustness and only 0.1 mm less accurate average distances from the human segmentations.

All segmenters, including ours, give clinically inadequate segmentation in some parts of some cases. Ours has the advantage that it almost always gives a good segmentation for most regions of an object (and frequently for all regions). Given that the user is led to notice that a portion of a segmentation is inadequate, e.g., by our outlier labeling described above or by display of the segmentation vs. the image data, the user then needs a means of editing the result. We plan to provide an editor consistent with our methodology, i.e., that produces a new m-rep that is a good segmentation. Our experience with landmark and contour data in the geometry-to-data match term used in our m-rep fitting to binary images suggests that a user giving a new contour or landmark in an image region that the m-rep from segmentation has missed should allow the program to modify the m-rep so that it fits well within the designated region.

5.4 What extensions of our method are in progress to make it more broadly useful?

Multi-figure objects and multi-object complexes. We believe it is useful to represent objects in multi-object complexes by separating the probability distributions on each object into a distribution on self-changes of the object and a distribution of the effect on it of its neighboring objects, or nearby parts thereof. It is possible that the success of our method on single objects modeled by a single medial sheet is partially due to our use of the m-rep object representation. We have some slight evidence that its ability to represent local twists, bends, and magnifications directly makes it, in combination with PGA, able to more stably extract principal directions and variances than PCA on spoke-end boundary points. However, we guess that in the applications discussed here, m-reps' strengths in representing inter-object relationships are more important than their

strengths in representing single objects. We have begun work to provide such estimations and segmentations of the bladder-prostate-rectum complex, with attractive initial results [Jeong et al., 2006].

Quantile functions and their regions. [Broadhurst, 2005b] shows how to extend the ideas of RIQFs to images of multiple derived texture features. In nonmedical applications this method of texture feature QFs have shown superior texture classification performance, and we expect that segmentation by our methods and these texture QFs from medical images in which texture is important, such as ultrasound, will be effective. Work to extend the description of objects' relation to bones using distance quantile functions is also showing attractive initial results. Extension of these methods to the multiple MRI intensities is also in progress.

In the texture work, QFs of texture features that are non-independent were statistically analyzed. The methods used there for extension of non-independent analyses to the intensity RIQFs for different local regions in our segmentation method has yielded improvements in training and application efficiency without losing segmentation accuracy.

The formation of regions of homogeneous mixture for DRIQF statistics is reported and evaluated in [Stough et al., 2007]. Stough is also investigating the possibility that DRIQF regions shift along an object boundary in such a way as to optimize the image match.

Bringing segmentations to the voxel scale. Voxel scale refinement may be useful when the image contrasts are high at object boundaries. [Saboo et al., 2006] reports an extension of atlas (mean) to object warps implied by m-reps to a space warp, followed by boundary landmark based diffeomorphism, followed by a refinement warp at the voxel scale by diffeomorphism methods. The production of within-object diffeomorphisms from an m-rep segmentation is still under study by Levy.

Prospective studies and studies on other new cases. The results reported are on retrospective, limited size image collections. Moreover, our method has been tuned by making improvements over kidney cases and more recently, the 5-patient within-patient male pelvis cases. We have recently begun to apply our method to such pelvis within-patient collections for other patients with slightly more challenging images, and initial results are comparable to the ones reported above. We are embarking upon studies on larger amounts of data and prospective data from the clinic.

5.5 Conclusions

We conclude

- Statistical training on geometry and on intensity patterns can yield human-rivaling segmentations in challenging 3D images.
- The better the statistical training, the better the performance.
- Initialization is still critical; landmarks or a few contours are useful for initialization, and initialization warps can provide an improvement over just similarity transforms.
- Region-based intensity pattern representations can be superior to voxel-by-voxel representations.
- Statistics on local region DRIQFs improves results over global DRIQFs.
- Training (fitting binary images) with models constrained to be geometrically legal is important.

Acknowledgements

We thank Xifeng Fang, Qiong Han, Derek Merck, Rohit Saboo, Christina Villarruel, and Graham Gash for method design, programming, and running programs; Sarang Joshi and J. Stephen Marron for contributions to the algorithms; Kevin Gorczowski, Manjari Rao, Sona Kalra for evaluations; and Julian Rosenman and Martin Styner for advice on methods, applications, and evaluations. The research on which this paper is based was partially supported by NIBIB grant P01 EB02779.

References

[Broadhurst et al., 2005a] R. E. Broadhurst, J. Stough, S. M. Pizer, and E.L. Chaney. Histogram statistics of local model-relative image regions. In *proc. of Int. Workshop on Deep Structure, Singularities and Computer Vision (DSSCV)*, pages 72-83, June 2005.

- [Broadhurst, 2005b] R. E. Broadhurst. Statistical estimation of histogram variation for texture classification. In *proc. of the Fourth International Workshop on Texture Analysis and Synthesis*, pages 25–30, Oct. 2005.
- [Broadhurst et al., 2006] R. E. Broadhurst, J. Stough, S. M. Pizer, and E. L. Chaney. A statistical appearance model based on intensity quantile histograms. In *Proc. of IEEE Int. Symposium on Biomedical Imaging*, pages 422–425, 2006.
- [Cates et al., 2006] J. Cates, P. T. Fletcher, and R. Whitaker. Entropy-based particle systems for shape correspondence. In *proc. of MICCAI Workshop Mathematical Foundations of Computational Anatomy*, pages 90–99, 2006.
- [Christensen et al., 1994] G. E. Christensen, R. D. Rabbit, and M. I. Miller. 3d brain mapping using a deformable neuroanatomy. *Phys. Med. Biol.*, 39:609–618, 1994.
- [Cootes et al., 1993] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam. The use of active shape models for locating structures in medical images. In *Proc. of Information Processing in Medical Imaging*, volume 687 of *Lecture Notes in Computer Science*, pages 33–47, 1993.
- [Cootes et al., 2001] T. F. Cootes and C. J. Taylor. Statistical models of appearance for medical image analysis and computer vision. In *Proc. of SPIE Medical Imaging*, volume 4322, pages 236–248, 2001.
- [Costa et al., 2007] M. J. Costa, H. Delingette, and N. Ayache. Automatic segmentation of the bladder using deformable models. In *proc. of IEEE Int. Symposium on Biomedical Imaging*, 2007.
- [Crouch et al., 2007] J. Crouch, S. M. Pizer, E. L. Chaney, Y. Hu, G. S. Mageras, and M. Zaider. Automated finite element analysis for deformable registration of prostate images. *IEEE Trans. on Medical Imaging*, in press.
- [Damon, 2007] J. Damon. Global geometry of regions and boundaries via skeletal and medial integrals. *Communications in Analysis and Geometry*, to appear.
- [Davies et al., 2007] R. Davies, C. Twining, T. Williams, and C. Taylor. Group-wise correspondence of surfaces using non-parametric regularization and shape images. In *proc. of Information Processing in Medical Imaging (IPMI)*, 2007.
- [Fletcher et al., 2004] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. on Medical Imaging*, 23(8):995–1005, Aug. 2004.
- [Foskey et al., 2005] M. Foskey, B. Davis, L. Goyal, S. Chang, E. L. Chaney, N. Strehl, S. Tomei, J. Rosenman, and S. Joshi. Large deformation 3D image registration in image-guided radiation therapy. *Phys. Med. Biol.*, 50, 2005.
- [Freedman et al., 2005] D. Freedman, R. J. Radke, T. Zhang, Y. Jeong, D. M. Lovelock, and G. T. Y. Chen. Model-based segmentation of medical imagery by matching distributions. *IEEE Trans. on Medical Imaging*, 24(3):281–292, 2005.
- [Han et al., 2007] Q. Han, D. Merck, J. Levy, C. Villarruel, E. Chaney, and S. M. Pizer. Geometrically proper models in statistical training. In *proc. of Information Processing in Medical Imaging (IPMI)*, 2007.
- [Hazlett et al., 2006] H. C. Hazlett, M. D. Poe, G. Gerig, R. Gimpel Smith, and J. Piven. Cortical gray and white brain tissue volume in adolescents and adults with autism. *Biological Psychiatry*, 59(1):1–96, Jan. 2006.
- [Jeong and Radke, 2006] Y. Jeong and R. Radke. Modeling inter- and intra-patient anatomical variation using a bilinear model. In *Proc. of Mathematical Methods in Biomedical Image Analysis (MMBIA)*, 2006.
- [Jeong et al., 2006] J. Jeong, S. M. Pizer, and S. Ray. Statistics on anatomic objects reflecting inter-object relations. In *proc. of MICCAI Workshop From Statistical Atlases to Personalized Models: Understanding Complex Diseases in Populations and Individuals*, pages 136–145, 2006.
- [Joshi et al., 2004] S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23(Suppl. 1):S151–S160, 2004.
- [Kass et al., 1988] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int. J. Computer Vision*, 1(4):321–331, 1988.
- [Kelemen et al., 1999] A. Kelemen, G. Szekely, and G. Gerig. Elastic model-based segmentation of 3-d neuroradiological data sets. *IEEE Trans. on Medical Imaging*, 18(10):828–839, 1999.

- [Leventon et al., 2000] M. Leventon, O. Faugeras, E. Grimson, and W. Wells. Level set based segmentation with intensity and curvature priors. In *Proc. of Mathematical Methods in Biomedical Image Analysis (MMBIA)*, pages 4–11, 2000.
- [Levina and Bickel, 2001] E. Levina and P. Bickel. The earth movers distance is the mallows distance: Some insights from statistics. In *proc. of the International Conference on Computer Vision (ICCV)*, pages 251–256, 2001.
- [Levy et al., 2007] J. H. Levy, R. E. Broadhurst, S. Ray, E. L. Chaney, and S. M. Pizer. Signaling local non-credibility in an automatic segmentation pipeline. In *proc. of SPIE Medical Imaging*, volume 6512, 2007.
- [McInerney and Terzopoulos, 1996] T. McInerney and D. Terzopoulos. Deformable models in medical images analysis: a survey. *Medical Image Analysis*, 1(2):91–108, 1996.
- [Merck et al., 2006] D. Merck, G. Tracton, S. M. Pizer, and S. Joshi. A Methodology for Constructing Geometric Priors and Likelihoods for Deformable Shape Models. *UNC Chapel Hill technical report*, <http://midag.cs.unc.edu>, 2006.
- [Montagnat and Delingette, 1997] J. Montagnat and H. Delingette. Volumetric medical images segmentation using shape constrained deformable models. In *proc. of CVRMed*, volume 1205 of *Lecture Notes in Computer Science*, pages 12–22, 1997.
- [Muller, 2007] K Muller, Y.-Y. Chi, J. Ahn, and J.S. Marron. High Dimension, Low Sample Size Principal Components; Estimating Eigenvalues of a Singular Wishart. In preparation for resubmission to *J. Amer. Stat. Assoc.*
- [Pizer et al., 2003] S. M. Pizer, P. T. Fletcher, S. Joshi, A. Thall, Z. Chen, Y. Fridman, D. Fritsch, G. Gash, J. Glotzer, M. Jiroutek, C. Lu, K. Muller, G. Tracton, P. Yushkevich, and E. Chaney. Deformable m-reps for 3d medical image segmentation. *Int. J. Computer Vision*, 55(2):85–106, 2003.
- [Pizer et al., 2006] S. M. Pizer, R. E. Broadhurst, J. Jeong, Q. Han, R. Saboo, J. Stough, G. Tracton, and E. L. Chaney. Intra-patient anatomic statistical models for adaptive radiotherapy. In *proc. of MICCAI Workshop From Statistical Atlases to Personalized Models: Understanding Complex Diseases in Populations and Individuals*, pages 43–46, 2006.
- [Pizer et al., 2007] S. M. Pizer, Q. Han, S. Joshi, P. T. Fletcher, P. A. Yushkevich and A. Thall. Chapter 8: Synthesis, deformation, and statistics of 3D objects via m-reps, in: K. Siddiqi and S. M. Pizer, *Medial representations: Mathematics, algorithms, and applications*, Springer, to appear in 2007.
- [Prastawa et al., 2003] M. Prastawa, E. Bullitt, S. Ho, and G. Gerig. Robust estimation for brain tumor segmentation. In *proc. of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 530–537, 2003.
- [Rao et al., 2005] M. Rao, J. Stough, Y. Y. Chi, K. Muller, G. Tracton, S. M. Pizer, and E. L. Chaney. Comparison of human and automatic segmentations of kidneys from ct images. *Int. J. Rad. Oncol. Bio. Phys.*, 61(3):954–960, 2005.
- [Ray et al.] S. Ray, J. Jeong, S. M. Pizer, K. Muller, and Q. Han. Sample size advantages of statistics on a nonlinear manifold to characterize nonlinear variation in a population. In draft.
- [Saboo et al., 2006] R. Saboo, J. G. Rosenman, and S. M. Pizer. GeoInterp: Contour interpolation with geodesic snakes. *UNC Chapel Hill technical report*, <http://midag.cs.unc.edu>, 2006.
- [Saboo et al, 2007] R. Saboo, C. Villarruel, E. L. Chaney, J. Rosenman, and S. M. Pizer. Segmentation of tubular objects in medical imaging by posterior optimization of m-reps. *UNC Chapel Hill technical report*, <http://midag.cs.unc.edu>, 2006.
- [Shen and Davatzikos., 2002] D. Shen and C. Davatzikos. HAMMER: Hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging*, 21(11):1421–1439, November 2002.
- [Stough et al., 2004] J. Stough, S. M. Pizer, E. L. Chaney, and M. Rao. Clustering on image boundary regions for deformable model segmentation. In *proc. of IEEE Int. Symposium on Biomedical Imaging*, pages 436–439, 2004.
- [Stough et al., 2007] J. Stough, R. E. Broadhurst, S. M. Pizer, and E. L. Chaney. Regional appearance in deformable model segmentation. In *proc. of Information Processing in Medical Imaging (IPMI)*, 2007.
- [Tsai et al., 2003] A. Tsai, W. Wells, C. Tempany, E. Grimson, A. Willsky. Coupled multi-shape model and mutual information for medical image segmentation. In *proc. of Information Processing in Medical Imaging*, volume 2732 of *Lecture Notes in Computer Science*, pages 185–197, 2003.

- [Yang et al., 2003] J. Yang,, L. H. Staib, and J. S. Duncan. Neighbor-constrained segmentation with 3d deformable models. In *proc. of Information Processing in Medical Imaging*, volume 2732 of *Lecture Notes in Computer Science*, pages 198-209, 2003.
- [Yushkevich et al., 2006] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig. User-guided 3d active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, 31:1116–1128, 2006.