Chapter 9
Statistical Applications with Deformable M-Reps
Anatomic Object Segmentation and Discrimination

Stephen Pizer, Martin Styner, Timothy Terriberry, Robert Broadhurst, Sarang Joshi, Edward Chaney, and P. Thomas Fletcher

Abstract There are many uses of the means of representing objects by discrete m-reps and of estimating probability distributions on them by extensions of linear statistical techniques to nonlinear manifolds describing the associated nonlinear transformations that were detailed in Chapter 8. Two important ones are described in this chapter: segmentation by posterior optimization and determining the significant shape distinctions that can be found in two different probability distributions on an m-rep with the same topology but from two different classes. Both uses require facing issues of probabilities on geometry at multiple levels of spatial scale. The segmentation problem requires the estimation of the probability of image intensity

S. Pizer
Medical Image Display & Analysis Group, University of North Carolina at Chapel Hill, USA,
e-mail: pizer@cs.unc.edu

M. Styner
Medical Image Display & Analysis Group, University of North Carolina at Chapel Hill, USA,
e-mail: styner@cs.unc.edu

T. Terriberry
Medical Image Display & Analysis Group, University of North Carolina at Chapel Hill, USA,
e-mail: tterribe@cs.unc.edu

R. Broadhurst
Medical Image Display & Analysis Group, University of North Carolina at Chapel Hill, USA,
e-mail: reb@cs.unc.edu

S. Joshi
Medical Image Display & Analysis Group, University of North Carolina at Chapel Hill, USA,
e-mail: sjoshi@cs.unc.edu

E. Chaney
Medical Image Display & Analysis Group, University of North Carolina at Chapel Hill, USA,
e-mail: edward_chaney@med.unc.edu

P.T. Fletcher
Department of Computer Science, University of Utah, USA,
e-mail: fletcher@sci.utah.edu

K. Siddiqi and S. Pizer (eds.) *Medial Representations – Mathematics, Algorithms and Applications*.

distributions given the object description; we describe a way of doing that by an extension of principal component analysis to regional intensity summaries produced using the object-relative coordinates provided by m-reps. Applications of both segmentation and determination of shape distinctions to anatomic objects in medical images are described. Also described is a variant on the segmentation program used in estimating the probability density on an m-rep; this program fits an m-rep to a binary image in a way that is intended to achieve correspondence of medial atoms across the training population.

9.1 Introduction and Statistical Formulation

Both segmentation, i.e., extraction, of objects from images and characterization of geometric differences between classes of objects are usefully accomplished in terms of deformable shape models. In segmentation a geometric model is deformed into the image data, allowing the method to reflect an understanding of what legitimate or typical shapes are. In characterizing the differences between shapes in two different populations, the differences are measured in terms of the deformation from one shape to another. Medial models provide a useful representation of the object or complex of objects that undergoes deformation and of the deformations themselves. Moreover, statistics on medial models are useful for both applications, specifying the typicality of a shape or the population of deformations between shapes in the two classes being compared. Finally, the segmentation application requires not only statistics on the geometry, i.e., on the medial models or their deformations, but also statistics on the image intensities, given a medial model. Because these intensities are best understood statistically in object-relative coordinates, the figural coordinates provided by m-reps are an important means of producing the image intensity statistics.

In Chapter 8 the geometry of discrete m-reps and statistics on these entities were discussed. This chapter discusses the use of these geometric representations and their statistics, as well as the statistics on image intensities in figural coordinates for segmentation of anatomic objects and object complexes. It also discusses the use of these geometric representations and their statistics for statistical shape difference characterization between classes of anatomic objects or object complexes extracted from medical images, e.g., between the hippocampi or lateral ventricles of healthy and schizophrenic individuals as extracted from magnetic resonance images.

In characterizing the difference between two anatomic populations the differences need not only to be specified statistically, but also this specification needs to include *where* the differences are and *what form of deformation* occurs there, for example, whether it is a local twist or a local bend or a local swelling or a local contraction. Also, in segmentation, a coarse-to-fine, i.e., successively more local approach has serious speed advantages for any given quality of segmentation. M-reps with their coordinate systems, their provision of multiscale statistics, and their

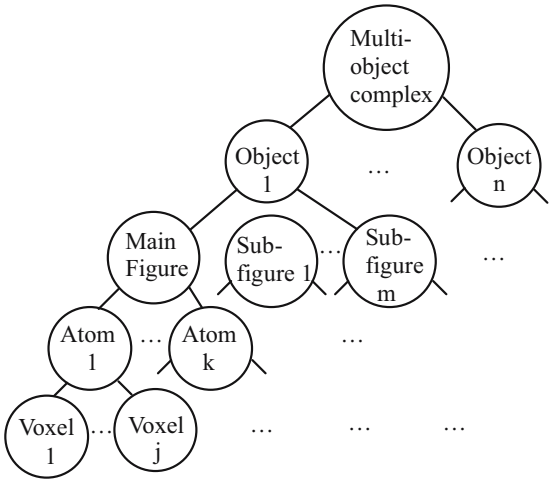


Fig. 9.1 A tree of objects, figures, medial atoms and voxels

medial basis' provision of both local width and local figural orientation are well matched to these needs.

More precisely, as illustrated in Fig. 9.1, consider a tree of geometrical entities such that the discrete m-rep at the root of the tree describes a whole object complex and such that the children of a node describe sub-entities which taken together make up that entity. For example, if the root node describes a complex of objects, its children would respectively describe each object making up the complex. Similarly, if a node describes an object made up of figures, its children would respectively describe the figures making up the object, their children might describe individual medial atoms, and their children might describe sequences of displacement on individual voxels. In each node is a collection of atoms made up of all its children, each atom with a value. The value of a node is the atom values of all the atoms making up that node. Then deforming the entity corresponding to a node deforms all of its sub-entities, and after that deformation we may move on to the sub-entities of the node and deform them further in some order. We refer to these stages at which processing occurs as *scale-levels.*

At each scale-level other than the top of the tree, an entity $\mathbf{m}$ has as set of neighbors $N(\mathbf{m})$, that are at nearby physical positions. It is useful to think of the probabilistic relationship among entities in terms of the value of each child of a node, given the value of that node and the conditional probability of a node given the values of its neighbors. The former describe inter-scale-level differences, and the latter describe inter-neighbor differences, i.e., differences across position. This view allows us to think of the problem with a Markov random field formulation in both the scale and positional dimensions.

That is, if the m-rep $\mathbf{n}$ is a child sub-entity of an m-rep $\mathbf{m}$, and $\mathbf{m} \rightarrow \mathbf{n}$ is the value that $\mathbf{n}$ takes as a result of the deformations of its ancestors and most recently as a result of its parent $\mathbf{m}$, we wish the conditional probability of the deformation describing the difference $(\mathbf{m} \rightarrow \mathbf{n}) \ominus \mathbf{n}$, given the parent node $\mathbf{m}$, where the symbol

$\ominus$ denotes the geodesic path between its two operands. Similarly, if $< N(\mathbf{n}) >$ describes the prediction of $\mathbf{n}$ based on its neighbors, we wish the conditional probability, $p(\mathbf{n}\ominus < N(\mathbf{n}) > |N(\mathbf{n}))$, of $\mathbf{n}\ominus < N(\mathbf{n}) >$ given the neighbor nodes $N(\mathbf{n})$. Because the essence of geometry is that entities are locally correlated, a thesis that for medial atoms of various anatomic objects is supported by our data, it is reasonable to condition $(\mathbf{m} \rightarrow \mathbf{n}) \ominus \mathbf{n}$, on only the parent node $\mathbf{m}$ and not on ancestors more distant in scale, and it is reasonable to condition $\mathbf{n}\ominus < N(\mathbf{n}) >$ on its immediate neighbors $N(\mathbf{n})$ and not on more distant entities.

In the work described here, we simplify the probabilistic formulation even further. We assume that $(\mathbf{m} \rightarrow \mathbf{n}) \ominus \mathbf{n}$ is statistically independent of $\mathbf{m}$ and that $p(\mathbf{n}\ominus < N(\mathbf{n}) > |N(\mathbf{n}))$ can be broken up into two factors, one describing the change independent of its neighbors and the other describing the interrelationship of it with its neighbors. Breaking things down according to this Markov formulation allows a segmentation or hypothesis test with final locality such that the total number of primitives at that level of locality is M (e.g., there are M voxels in the objects being segmented or at which shape differences are being tested) to operate in $\mathcal{O}(M)$ time rather than the $\mathcal{O}(M^2)$ that are required when the relation of every primitive with every other one must be dealt with.

The geodesic differences between m-reps used in the foregoing formulation are in the same symmetric space as the subtrahend and the minuend. That is, the, the geodesic differences of a collection of medial atoms is the collection of differences of the corresponding atoms, and the difference of two atoms is the Cartesian product of the corresponding components, as illustrated by the difference between interior slab atoms in the following:

1. The difference of the hub positions, which like a hub position itself is a vector in $\mathcal{R}^n$.
2. The "difference" of the spoke lengths, which is the ratio of these lengths giving the magnification of one into the other, and thus like a length itself is a scalar in $\mathcal{R}^+$.
3. The "difference" of each spoke position on the unit sphere S^2 with the corresponding spoke's position on S^2, which can be understood as a position on S^2. There are difficulties with differences of angle differences associated with having to specify a reference angle; these will not be further discussed here.

As a result, statistics on such geodesic differences can be accomplished by the same methods of computing means and principal geodesics described in Chapter 8.

Finally, consider the probabilities on differences of m-reps that are the target of statistical characterization of inter-class differences. These differences of m-reps are again in the same symmetric space as the subtrahends and minuends. One requires methods of hypothesis testing that yield the significance of distinctions in probability distributions in this symmetric space and, as well, the location of such significant changes, for various levels of locality.

In Section 9.2 we introduce segmentation via posterior optimization of deformable m-reps with an overview of the approach. We find that two log probability densities are needed, one measuring the geometric typicality of an m-rep and the

other measuring the match between the m-rep and an image. In Section 9.3 we discuss how to train the first probability density, given binary images of sample objects, and how to measure this geometric typicality on any m-rep, given this training. In Section 9.4 we discuss estimating the probability density on image intensities given a medial model and how to measure this probability density on any target image. In Section 9.5 we conclude our discussion of segmentation by specifying the segmentation scale at the smallest scale level, that of the voxel, followed by the excellent results obtained using our multi-scale method using the geometric and intensity probabilities. In Section 9.6 we discuss means of hypothesis testing based on m-reps for statistical characterization of shape differences between populations of objects or object complexes. Section 9.7 gives some examples of results using this method. In Section 9.8 we discuss the apparent strengths and weaknesses of the medial methods we propose for the segmentation application and characterization of shape differences application, as compared to alternative object representations. In that section we also discuss work that remains in both these methods of application of m-reps and in the formulation of m-reps themselves and their statistics.

9.2 Segmentation by Posterior Optimization of Deformable M-Reps: Overview

Published studies by others and our own research results strongly suggest that segmentation of a normal or near-normal object (or objects) from 3D medical images in all but the simplest cases will be most successful if it uses (1) knowledge of the geometry of not only the target anatomic object but also the complex of objects providing context for the target object and (2) knowledge of the image intensities to be expected relative to the geometry of the target and contextual objects.

We use the general segmentation approach already shown by others to lead to success ((Cootes et al., 1993; Staib and Duncan, 1996; Delingette, 1999), among others; also see (McInerny and Terzopoulos, 1996) for a survey of active surfaces methods), namely deforming a geometric model by optimizing an objective function that includes a geometry-to-image match term which is constrained by or summed with a geometric typicality term. In this approach a model of the object(s) to be segmented is placed in the target image data and undergoes a series of transformations that deform the model to closely match the target object.

In computer vision an important class of methods uses explicit geometric models in a Bayesian statistical framework to provide *a priori* information used in posterior optimization to match the deformable shape models against a target image. Using this approach, we start from a statement of the segmentation objective as finding the most probable conformation of the target object(s) $\mathbf{m}$ given the image I, i.e., of computing $\text{argmax}_{\mathbf{m}} p(\mathbf{m}|I)$. Here $\mathbf{m}$ is the geometric representation of the target object(s), in our case the tree of medial atom meshes that comprises an m-rep, and I is a tuple formed by a 3D array of image intensities. The probability density $p(\mathbf{m}|I)$

is frequently called the *posterior density*, so the method is called one of *posterior optimization* (Duda et al., 2001).

By Bayes rule, $\text{argmax}_{\mathbf{m}} p(\mathbf{m}|I) = \text{argmax}_{\mathbf{m}}[\log p(\mathbf{m}) + \log p(I|\mathbf{m})]$. Thus the geometric typicality term ideally measures the logarithm of the so-called *prior* probability density, the probability density that the candidate geometric entity exists in the population of objects, as described in Chapter 8. And the geometry-to-image match term ideally measures the logarithm of the so-called *likelihood*, the probability density that the target image values, relative to the candidate geometry, would arise in the population of images from that modality. As a fundamental means of obtaining efficiency, we optimize such an objective function for successively smaller spatial tolerances (spatial scales), where each of the spatial scale levels are object-relevant: the object complex, the object, the slab (or tube) figure, the figural section, and the voxels not only interior to the object(s) but also the voxels between them, which we call *interstitial* voxels.

The success of the deformable shape models posterior optimization approach depends on the object representation, i.e., the structural details and parameter set for the deformed model, as well as on the form of the objective function. The most common geometric representation in the literature of segmentation by deformable models is made up of directly recorded boundary locations, sometimes called *b-reps* (Cootes et al., 1993; Kelemen et al., 1999), also see papers surveyed by (McInerny and Terzopoulos, 1996). Our m-reps representation (Fig. 9.2), principal geodesic analysis to produce its statistics, and the associated segmentation method use a medial representation intended to produce improved and/or more efficient segmentations for the reasons given in Chapter 8, Section 11. The most relevant of these advantages for this application are the efficient training of the prior it provides, its ability to provide a coordinate system in which to describe intensities probabilistically, and its inherent multi-object, multi-scale nature, which leads to effectiveness and efficiency of segmentation of single or multiple objects. However, small indentations and protrusions of anatomic objects are impractical to model medially. Our approach to solving this problem is to implement a non-medial voxel stage described in Section 9.5.1.

M-reps, combined with the voxel-level representation, provide their advantages over other deformable object representations at the expense of a level of complexity that required the development of special theoretical underpinnings, software,

Fig. 9.2 M-rep modeled kidney with its medial mesh, a liver model that is made from two figures, one for each lobe, and male pelvis model made from multiple objects (two bones, bladder, rectum, prostate). The kidney model also shows the underlying representation of a sampled medial surface and a tiled boundary

and validations. Largely automatic segmentation by large to small application of deformable m-reps has been implemented in software called *Pablo* (Pizer et al., 2005b) that accomplishes 3D segmentations in a few minutes. Software for building and training models has also been developed. The methods underlying this software and its abilities are the subject of Sections 9.2–9.5.

The next two sections give a more specific picture of Pablo's method (Section 9.2.1) and operation (Section 9.2.2).

9.2.1 Segmentation Method: Posterior Optimization for Multiscale Deformation of Figurally Based Models

Our method for deforming a model into image data typically begins with a manually chosen initial positioning of the mean model, frequently via choosing a few rough landmark positions. The segmentation process then follows a number of stages of segmentation at successively smaller levels of scale. The spatial tolerance of the resulting segmentation can be large at the largest scale level but decreases as the scale gets smaller.

As illustrated in Fig. 9.3, at each scale level, i.e., level of the tree shown in Fig. 9.1, the same log prior + log likelihood objective function is optimized by geometrically transforming the entities at that scale level, using a transformation

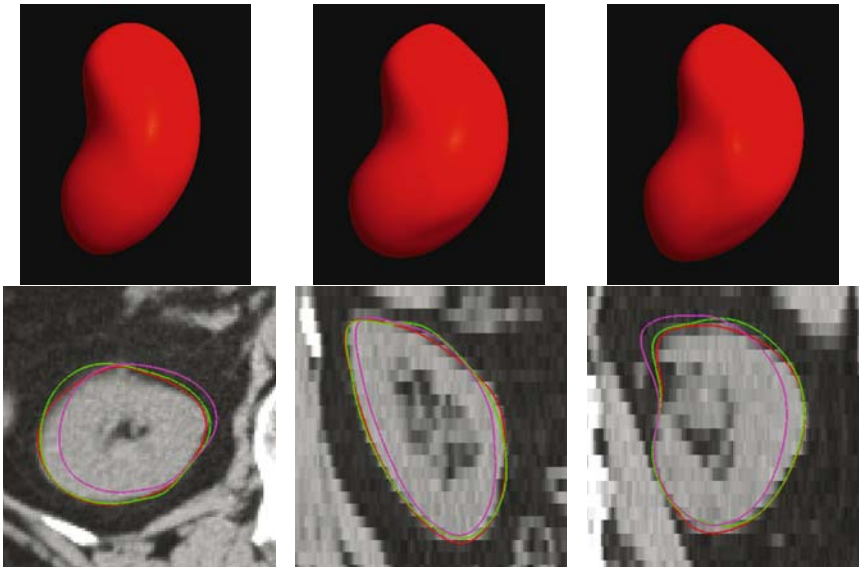


Fig. 9.3 Stage by stage progress of deformable m-rep segmentation of the kidney. Top: rendered 3D view, after model alignment via landmarks, the figure stage, and the figural section (atom) stage. Bottom: results on axial, sagittal and coronal CT slices. Each image compares progress through consecutive stages via overlaid curves: magenta—aligned position; green—post object stage; red—post atom stage

global to the respective entity. Thus, at the largest scale level, the object ensemble stage, the whole object ensemble undergoes a global transformation. At the next smaller scale level, each object making up the object ensemble separately undergoes a transformation global to it. And as the computation moves to successively smaller scale stages, successively smaller entities making up the entities at the next larger scale level, namely figures, subfigures, and medial atoms, are optimized with a transformation global to each of them. The series of optimizations concludes with a small relocation of all of the voxels in the image being optimized.

At all of these scale levels, we follow the strategy of iterative conditional modes, so the algorithm cycles among the component entities in random order until the group converges.[1] For example at the figural atom stage, the algorithm cycles through the atoms in random order.

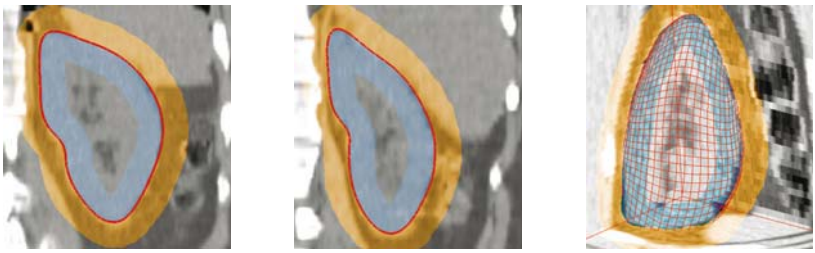
At each scale level larger than the voxel scale level, the geometric transformation of the entity is made up of a typically deterministic similarity transformation and a maximum posterior warp. The similarity transform, a translation, rotation, and uniform magnification, aligns the entity to neighboring entities of the same type (objects to neighboring objects, medial atoms to neighboring atoms), except it aligns to landmarks at the largest scale. The warp is formed from a few principal geodesics (see Chapter 8) of the deformations of that entity experienced in the training data. At the voxel scale level, the optimization is over displacements per voxel of only a few voxel widths. The result is that we typically optimize 6 or fewer parameters per entity, providing efficiency and convergence of the segmentation at that scale level.

At each scale level we use the conjugate gradient method to optimize the log prior + log likelihood objective function. The log prior metric is detailed further in Section 9.3. As detailed in Section 9.4.2, we have implemented a way of computing the log likelihood that measures the geometry-to-image match based on probability densities on intensity distribution features in various figural-coordinate-specified regions inside and outside of the object (Fig. 9.4) such that each region is expected to be a constant mixture of tissue types (Broadhurst et al., 2005).

9.2.2 Segmentation Method: User Operation

M-rep-defined objects can be viewed as a boundary mesh (at any of a number of vertex spacing levels), a rendered surface, a collection of points at the aforementioned boundary vertices, or a medial atom mesh. Most users find the first two of these the most useful. Images are normally viewed in a tri-orthogonal display, with the three possible slice directions fixed to the cardinal within-image and cross-image slice directions given by the stored target image. The displayed object can be presented together with the intensity display (see Fig. 9.3). Moreover, we also provide

[1] The convergence properties are shared with all iterative conditional modes methods and are based on the underlying Markov random field. In practice, convergence always occurs, but sometimes the convergence is to a local maximum of the objective function rather than the desired global maximum.



a b c

Fig. 9.4 Boundary-relative regions used for measuring geometry-to-image match to a kidney. (**a**, **b**) Example from two different patients displayed in 2D cuts. The kidney interior region is portrayed in blue, and the kidney exterior region is portrayed in orange. (**c**) A mesh showing in 3D the m-rep implied boundary of the kidney, and the kidney interior and exterior regions in two orthogonal cuts through the 3D image

a boundary display mode on the displayed slice, in which the 3D object does not appear but the curve of its intersection with the displayed slice(s) is displayed on that slice (those slices).

Using these viewing mechanisms, the user either chooses the location of preselected landmarks in the target image, which is then used as the basis of an Procrustes initialization of the model, or he or she manually initializes the chosen model by placing it in an initial position relative to the 3D image (for example, see Figs. 9.3-bottom row, 9.4c, and 9.9-bottom left). The initialization transform derived from the landmarks is frequently a similarity transform, but we have found it also useful that this landmark-based transform optimize in the shape space of the principal geodesics of the object with a data-match term given by the sum of squared model landmark to image landmark squared distances, with each squared distance divided by its tolerance squared.

The landmarks on the model are chosen as a specified spoke end. These landmarks appear as colored spots on the base model in the display space. These landmarks can also be used for editing an m-rep in the middle of the optimization process or as another term in the geometry-to-image match.

The user is also given control of the values of the weights controlling the strength of the geometric typicality term in the objective function, relative to the geometry-to-image match term. However, since the two terms are now both Mahalanobis distances, the default weight of unity needs seldom be changed.

9.3 Training and Measuring Statistical Geometric Typicality

To be able to measure a log prior, one needs a parametrized function that one can evaluate with any m-rep for the desired object as the argument. Section 9.3.1 describes the means for training the parameters of this prior probability density on m-reps that is then used to measure geometric typicality of any candidate m-rep appearing in the optimization of the log posterior. This training of the prior is done

by principal geodesic analysis of m-reps fit to binary images extracted from training greyscale images. Section 9.3.1 describes both the fitting of m-reps to binary images and how principal geodesic analysis is used at multiple scales to produce the prior probabilities needed for the various scale levels. Section 9.3.2 describes the means for measuring the log prior at multiple scales needed in the multiscale segmentation procedure.

9.3.1 M-Rep Model Fitting and Geometric Statistics Formation

Model-building must designate the figures making up an object or object ensemble, give the size of the mesh of each figure, and give the way the figures are related. It must also specify each medial atom in the model forming the mean object or object ensemble and the variability of these at many scale levels. Illustrated in the panels of Fig. 9.5 are m-rep models of a variety of anatomic structures that we have built.

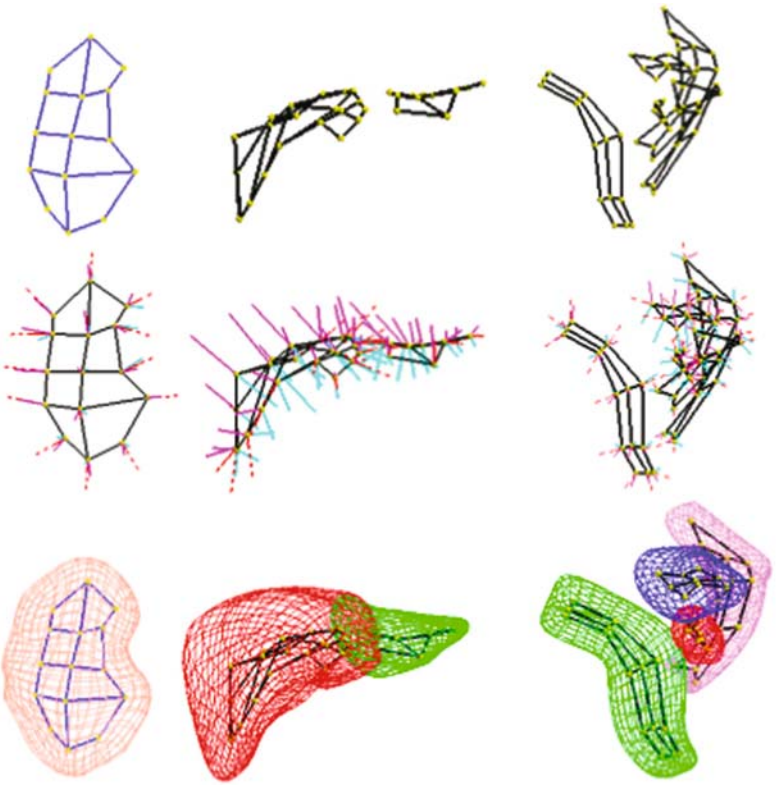


Fig. 9.5 M-reps for a kidney, a liver, and a male pelvis. *Top row*: mesh of atom hubs; *middle row*; mesh of medial atoms (including spokes); *bottom row*: the implied boundaries shown with atom mesh(es)

In the following we sketch our model building procedure, leaving the details of how we meet this challenging goal to other papers (Merck et al., 2006).

Because an m-rep is intended to allow the representation of a whole population of an anatomic object across patients, we build it based on a significant sample of instances of the object. Typically we use some tens of instances, say 50. For each instance we begin with both a 3D binary image representing the interior of the object, typically manually segmented, and an associated 3D greyscale image (CT or MRI or another modality).

Styner et al. (2003a) describe a tool for producing m-rep models from such binary image samples, based on the principle that effective segmentation depends on building a model that can easily deform into any instance of the object that can appear in a target image. We can use this tool to compute the set of appropriate figures at a given level of approximation from a training population, or we can choose the figures based on anatomic expertise to correspond to named anatomic structures. The tool measures the level of approximation in the figure computation step via error in volume overlap (typically 98%). In either case, given the figures, the tool chooses the number of atoms in each figure as the minimal number that can fit every training instance to a given error measured by the mean absolute distance of the surfaces (typically 5% of the average radius).

More recently we have completed a stable web-sharable tool called *Binary Pablo* for fitting an m-rep model to each member of a collection of binary images and deriving the Fréchet mean and principal geodesic modes and variances (Merck et al., 2006). Once a base model is generated, we use Pablo to deform it into the binary segmented training images. The program optimizes an objective function that has an "image match" term giving an average distance between the boundary implied by the m-rep and the binary image boundary, and three geometry terms: (1) giving an average squared-distance between each atom and the geodesic average of its neighbors, thus producing a regular mesh of atoms; (2) discouraging folded objects by penalizing rS_{rad} eigenvalues $\varepsilon 1$ (see Chapter 3); (3) giving a squared-distance from a reference m-rep. The sharable version only operates for single-figure objects, but versions that fit m-reps to multi-figure objects and to multi-object complexes are available in our research toolkit.

Given the m-rep models for all the training cases (Fig. 9.6), we use a tool initially developed in Dam et al. (2004) and further developed by Lu (Lu et al., 2003) to compute the mean model and the principal-standard-deviation-weighted principal geodesics describing its variability. This tool uses the method of principal geodesic statistics on symmetric spaces described in Chapter 8, Section 7. As with linear statistics, each principal geodesic has an associated variance, and moving along that geodesic gives a principal mode of variation of the population of m-reps.

The statistics at one scale level need to describe the variability of the geometric entity at that scale level after the variability at the larger scale levels has been accounted for and after alignment to neighboring entities has been done. Description of this residue statistics, based on the theory of Markov random fields, is given in Lu et al. (2003).

Fig. 9.6 *Left*: A subset of our population of training kidneys. *Right*: the mean of the population and the mean $+/-1$ standard deviation in each of the first two principal geodesic modes

With these means and a number of principal geodesics chosen to capture some fraction of the variance at that scale level, deforming a geometric entity at that scale level involves aligning the object to its neighbors and then computing the coefficients of the principal geodesics of the deformation of that entity.

9.3.2 Measuring Statistical Geometric Typicality

The geometric typicality that we wish to use is $\log(p(\mathbf{m}))$, or in the case that $\mathbf{m}$ has neighbors $N(\mathbf{m})$, $\log(p(\mathbf{m}|N(\mathbf{m}))$. But except for an additive constant and a constant multiplier of -0.5, when the principal geodesic analysis given in Section 8.7 is used, the log probability density in the symmetric space at any scale level is just a Mahalanobis distance in a tangent space to that symmetric space. Thus, when optimizing in the space of principal geodesics, we are optimizing over the weights a_i of the projections $\mathbf{v}^i$ of the unit-variance principal geodesics onto the feature space tangent plane at the mean. For any value of these a_i, and given the variances σ_i^2 of the principal geodesics in that tangent plane that are derived in the principal geodesic analysis, the Mahalanobis distance of $-\sum_i a_i^2$ forms the geometric typicality measure.

As discussed earlier, at all scale levels but the global one this geometric typicality metric of the relevant geometric entity needs to reflect its shape properties but also its relation to immediately neighboring peer entities. This can be accomplished with principal geodesics that were computed with augmenting atoms in adjacent objects or figures (see Chapter 8, Section 7).

Two special neighbor relations deserve comment. One is the non-interpenetration relation among very nearby (possibly abutting) objects (see the male pelvis in Fig. 9.2). Not only the correct relative position, orientation and size need to be reflected in the geometric typicality, but also an interpenetration of the figures needs to result in a low geometric typicality. The second neighbor relation of note is that between a protrusion or indentation *subfigure* to the "host" figure on or into which it sits (see liver in Figs. 9.2 and 9.5) or the relation between an object and a nearby, possibly abutting, object. In Chapter 8 we argued that the subfigure should ride on the boundary implied by the host's representation and be known in the figure-relative

coordinates of the host. The augmentation idea mentioned as applying to nearby objects uses a similar concept. Thereby we can make measurements of typicality in terms of the position of the subfigure (or related object) relative to the host, the orientation of the subfigure relative to the host, and the size of the subfigure relative to the host. When slight modifications of the hinge atom relationship are created due to motions in symmetric spaces not maintaining the relationship of hinge atoms to their host figure boundary, we find success in simply projecting the hinge atoms back onto the host boundary along host surface normals (interpolated spokes).

9.4 Training and Measuring Statistical Geometry-to-Image Match

Methods for training and measuring a probability density on image intensities must do so in a way respecting correspondence of locations across the population. There is much good work on correspondence, e.g., (Davies et al., 2002; Yushkevich et al., 2005), but here we suggest that correspondence be obtained through object-relative coordinates (Fig. 9.7). For m-reps that means that the figural coordinates provided by $\mathbf{u} = (u, v, \phi, \tau)$ within figures (see Chapter 8, Section 3) and by $\mathbf{u} = (v, w, \phi, \tau)$, within inter-figural blend regions (see Chapter 8, Section 8) provide the means of correspondence. More precisely, intensity statistics are done with respect to $I(\mathbf{u})$.

Recall that within an object main figure and within a subfigure outside of the blend region, (u, v) measures relative location along the medial sheet, ϕ expresses which side of the medial sheet the location is or at the end where in the transition between the sides the location is, and τ gives the fraction of the distance along the spoke from the medial end to the boundary end. For interfigural blend regions between a subfigure and a host figure v and ϕ are the cross-figure figural coordinates of the subfigure and $w \in [-1, 1]$ moves along the blend from the curve on the subfigure terminating the blend $(w = -1)$ to the curve on the host figure terminating the blend $(w = +1)$. Section 9.4.1 describes the computation transforming between Euclidean coordinates $\mathbf{x}$ and figural coordinates $\mathbf{u}$. Between objects one must interpolate between the figural coordinates of the nearby objects. The means

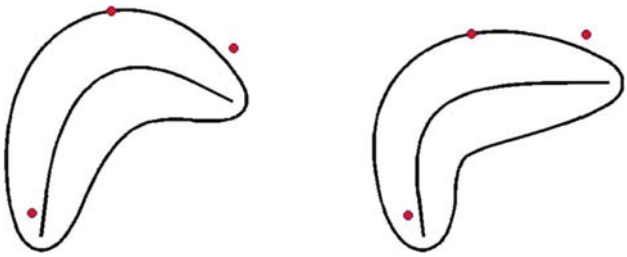


Fig. 9.7 Correspondence over deformation via figural correspondence. In each pair of corresponding marked points, the two points have the same value of the figural coordinates $\mathbf{u} = (u, v, \phi, \tau)$

of this interpolation is still a subject of research, but one of the options is described in Section 9.5.1.

We have used two basic methods that go from m-reps and associated greyscale images to geometry-to-image match functions on an image given an m-rep. The method we used first (Stough et al., 2004) was based, like that of the active shape method of Cootes et al. (1993), on normalized correlation between cross boundary intensity profiles and template profiles determined in training. However, in our method the template in each profile was chosen from among a limited number of possibilities chosen by clustering profiles during training, and values needed in the normalizations of the target profiles at each boundary vertex were also determined during training, thus stabilizing the normalization. Both normalized correlation methods produce a log probability density only under the poor assumption that the profiles are uncorrelated and that the tissue mixture at a voxel in the template can be expected to be precisely the same as that in the corresponding voxel in the target image. To overcome the first weakness Ho (2004) argues for improvements based on multiscale profiles, produced by a variant on Gaussian weighting across but not along the profiles.

Either variant of this profile match method can be expected to achieve less success than our new method, which is designed to produce log probabilities without these faulty assumptions. Our experiments on kidney segmentation, sketched in Section 9.5.2 and detailed in (Broadhurst et al., 2006), showed the new method to give better results in practice. Thus we describe only the new method, which generates a log likelihood on discrete quantile functions from the intensities in regions relative to the m-rep. It is detailed in Section 9.4.2.

9.4.1 Transforming Between Figural and Euclidean Coordinates

The geometry-to-image match term in the objective function requires object-relative image positions $\mathbf{x}(\mathbf{u})$ to be computed in large number. Thus, interpolation within $I(\mathbf{x})$ must be very efficient. In Pablo at present, this transformation $\mathbf{x}(\mathbf{u})$ is done through the mechanism of subdivision surface methods (Thall, 2004), as described below. Han is developing a more accurate method based on the interpolation of medial atoms (see Chapter 8) and is seeing how to make it adequately speedy.

In the subdivision surface method we interpolate the boundary first and consequently can interpolate medial atoms at any position on the sheet of atoms. The implied boundary is computed from the set of atom spokes connected into quadrilateral and triangular tiles both within figures and in interfigural blend regions (Figs. 8.3 and 8.13). The boundary interpolation is accomplished by a variation of the very efficient Catmull-Clark subdivision (Catmull and Clark, 1978) of the mesh of polygonal tiles. Thall's variation (Thall, 2004) of Catmull-Clark subdivision produces a limit surface that iteratively approaches a surface interpolating in position to spoke ends and with a normal interpolating the respective spokes. That surface is a B-spline at all but finitely many points on the surface. The program gives control of a tolerance on the normal and on the closeness of the interpolations.

The resulting B-spline allows the computation of both boundary positions **b** and boundary normals **U**, which are spoke directions there. Interpolating the medial radius r as well as u and v at such boundary positions allows the computation of $\mathbf{x}(\mathbf{u}) = \mathbf{b} + (\tau - 1)\,\mathbf{U}$.

Points **x** can also be given a figural coordinate **u** by finding the figural coordinates of the closest medially implied boundary point, using the boundary normal or the gradient of the distance function as the spoke direction, and calculating τ from the intersection of this spoke with the sheet of hubs. This calculation, however, is fraught with danger, since the boundary may be inadequately smooth.

9.4.2 Geometry-to-Image Match via Statistics on Discrete Regional Quantile Functions

9.4.2.1 Conceptual Basis for Statistics on Intensity Quantile Functions

Any efficient geometric description does not capture all there is to say about the biology or physics of the individual being modeled. Thus for a given medially specified object or complex of objects, the variation between different images of the same class of object frequently is due not only to intensity noise but more so to the variation of the materials of which the object are made and of the variation across the object of the weights of those materials making up the materials mixture. Thus in medical images there is variation across patients of the locations of specified tissue types within and between their respective objects. This suggests that point-by-point correspondence as provided, for example, in the active shape models and active appearance models of (Cootes et al., 1993, 1999), where the probability densities are on corresponding intensity values, be replaced by probability densities on regional collections of intensities, ignoring the particular locational correspondences within these regions. In particular, this suggests probabilities on intensity summaries, such as histograms, of regions expected to have uniform mixtures of tissue types.

Our regional intensity summary based match method (Broadhurst et al., 2005; Pizer et al., 2005a) uses a region inside the object and a region outside the object (Fig. 9.4) and sometimes subregions of these regions defined according to figural geometry.

The feature space formed by using the bin counts of histograms of intensity provides a poor basis for probabilistic analysis. The weakness is exemplified by the fact that the average of two unimodal histograms in this form would be a bimodal histogram, rather than a unimodal histogram whose center is between the two being averaged. In the following we argue that instead representing the regional intensity collection by the curve of intensity values versus quantile (regional intensity quantile function, or *RIQF*) allows an effective log probability density to be calculated by factor analysis. Also, histogram bin counts as features suffer from quantization effects, i.e., binning errors, while discrete RIQFs do not since no arbitrary bin boundaries are selected.

The RIQF of an intensity distribution i can be shown to be the inverse of the cumulative distribution function I of i. Discretely sampling the RIQF yields the *discrete RIQF (DRIQF)*. The DRIQF is an n bin quantile function where each bin j, representing $1/n$ of the probability distribution area, stores its average image intensity i_j. Considering these values in sorted order, the DRIQF for region k can be represented as a vector $i^k = (i_1^k, \ldots, i_n^k)$. Computing this vector requires partial sorting of the list of N intensities in the region, taking $O(Nlog(n))$ time.

The effectiveness of using standard statistical tools to construct a probability distribution of RIQFs depends on the fact that the space of RIQFs has several known linear properties related to Euclidean distance and thus mean and principal components. Euclidean distance between RIQFs corresponds to the Mallows distance (Mallows, 1972; Levina and Bickel, 2001) between the corresponding probability distributions, defined as follows. For two continuous one-dimensional distributions with cumulative distribution functions Q and R, and RIQFs $q = Q^{-1}$ and $r = R^{-1}$, respectively, the Mallows distance between them is defined as the Minkowski L_2 norm between q and r:

$$M_2(q,r) = \left(\int_0^1 |Q^{-1}(t) - R^{-1}(t)|^2 dt \right)^{1/2} = \left(\int_0^1 |q(t) - r(t)|^2 dt \right)^{1/2}.$$

The Mallows distance can be shown to measure the work required to change one distribution into another by moving probability mass, i.e., the Earth Mover's distance between the corresponding probability distributions, intuitively a good measure of difference between RIQFs. For DRIQFs q and r, the Mallows distance is defined as the L_2 norm of the vector difference between q and r:

$$M_2(q,r) = \left(\frac{1}{n} \sum_{j=1}^{n} |q_j - r_j|^2 \right)^{1/2}.$$

Location and scale changes to any probability distribution, or changes in any affine combination of the DRIQF values, are linear in the space of DRIQFs. Several families of common continuous distributions, including Gaussian, uniform, and exponential distributions, are parameterized by location and scale parameters. Thus, DRIQFs of each of these families of distributions exist in a two-dimensional linear subspace. Also, the Euclidean average (or any linear combination) of a set of DRIQFs from one of these families of distributions results in a DRIQF contained within the family and having means and standard deviations averaging (or correspondingly linearly combining) the respective means and standard deviations. For example, the Mallows distance between two Gaussian distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ is $\sqrt{(\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2}$. The average in this space of a set of RIQFs corresponding to Gaussian probability densities is a Gaussian with a mean and standard deviation equal to the average mean and standard deviation of the set of Gaussians. However, a weakness of the space is that for probability distributions composed of a mixture of multiple underlying unimodal distributions, changing the mixture amount is a nonlinear operation.

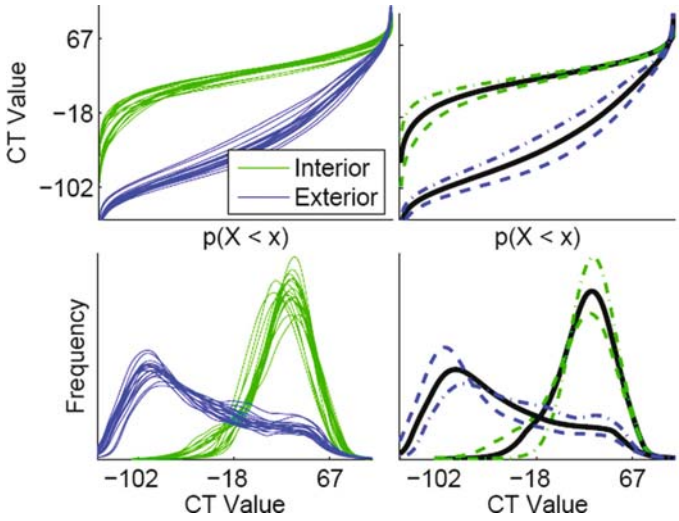


Fig. 9.8 Bladder DRIQFs (*top*) and corresponding histograms (*bottom*). *Left*: training samples; *right*: learned mean and ± 2 standard deviations along the first principal direction

The consequence of the foregoing is that analysis of regional intensity distributions can be captured by linear statistics on their DQRIFs, which can efficiently capture variation in location and scale change. The method is not effective for dealing with multimodal probability distributions with widely separated peaks and varying interpeak separations, but our experience is that it works well for unimodal probability distributions and even multimodal probability distributions whose peaks are not widely separated.

DRIQFs of interior and exterior regions of the bladder in 15 CT images are shown in Fig. 9.8. The first two principal directions of variation of the interior and exterior regions capture 95% and 97% of the variation, respectively. DRIQFs of subregions can also be constructed; this example and this discussion are only in terms of interior and exterior regions. In this example, the contribution of each voxel to the DRIQF is Gaussian weighted by its distance to the surface. This allows narrow regions to be defined that have larger capture ranges and smoother objective functions during segmentation than equivalent non-weighted regions. In each region, gas and bone intensities have been automatically removed from the distribution using a threshold, and the probability of each is independently measured. The Mallows distance is sensitive to the variation in these intensities due to their extreme intensity values compared to the differences in fat and tissue intensities.

9.4.2.2 Training Probability Densities on Regional Intensity Quantile Functions

The probability densities on DRIQFs that we use are estimated by principal component analysis of the DRIQFs, taken as feature vectors. Then the geometry-to-image

match of the DRIQF obtained from a target image region is a Mahalanobis distance based on this principal component analysis. In the following we detail the estimation of this Mahalanobis distance function.

We model the variation in our DRIQF feature space as a multivariate Gaussian distribution. The dimension of this space is equal to the number of bins used to represent a DRIQF, which is typically 200. This often results in a high dimension, low sample size situation, which prevents us from estimating a full rank multivariate Gaussian model. Therefore, we use principal component analysis to estimate a low dimensional subspace, typically of dimension 2–4, in which we build a Gaussian model. We then measure the expected distance to this subspace by summing the remaining eigenvalues since during segmentation we expect to estimate the probability of many image regions that are not typical of the training regions. Thus, the final Gaussian model for each region is of dimension equal to the number of eigenmodes plus one.

The geometry-to-image match function is the resulting Mahalanobis distance function. Intuitively, the Mahalanobis distance of a target DRIQF is equal to the Mallows distance between the probability distribution corresponding to the target DRIQF and that corresponding to the mean RIQF, modified by the standard deviations in each direction of the Gaussian model. Thus the Mahalanobis distance is a natural enhancement of the Mallows distance that accounts for the variability in the training set.

The training data on which the principal component analysis is done is formed as follows. For each training case we have a greyscale image, a binary image, and an m-rep fit to the binary image as discussed in Section 9.3.1. Voxel correspondences specified by m-rep based figural coordinates (Section 9.4.1) allows us to compute the set of DRIQFs for each object-relative image region across the training images. When determining if a voxel belongs in a region, we initially use the binary image, not the m-rep, to label voxels as being inside or outside the object. This allows us to define mean DRIQFs that correctly provide references for the Mahalanobis distances used to form the geometry-to-image match. These DRIQFs do not, however, measure the expected variation of the actual training segmentations. Therefore, we estimate the covariance of the DRIQF in each region from the DRIQF values based on m-rep region labeling minus the already computed respective mean DRIQF, which is based on binary labeling.

9.5 Pablo Details and Results

9.5.1 The Voxel-Scale Stage of Segmentation

After all of the stages of segmentation that modify the medial atoms, an m-rep has undergone transformation from the beginning model (typically the mean of the global object complex or object). Figural coordinates allow this transformation to

be understood as a diffeomorphism within all of the objects making up the complex represented by the m-rep. This warp can be interpolated into a chosen portion of 3-space including the complex, including the interstitial space between multiple objects or figures, if the complex is made up of more than one figure. A further finer scale transformation on that portion of 3-space can then be determined.

We interpolate the transformation from the objects to the surrounding 3-space, as follows. Each implied boundary position of the m-rep is understood as the tip of a particular m-rep spoke, either one of the basic representation or one interpolated from it. That spoke is from a medial atom $\mathbf{m}(1)$ at particular figural coordinates that allow it to be associated with a corresponding atom $\mathbf{m}(0)$ in the original m-rep model. Paths $\mathbf{m}(t)$, $0 \leq t \leq 1$, in the abstract space of atoms between the original atoms and the corresponding transformed atoms can be found according the mathematics in Chapter 3, Section 3.3, such that at every t the m-rep is unfolded and thus the continuous transformation of m-rep interiors is diffeomorphic. These paths can be sampled in t to produce a path of the corresponding spoke ends, and this sequence of positions can be used as a boundary condition in a landmark interpolation method. For example, one can use the thin-plate spline interpolation (Bookstein, 1991) on each of the corresponding successive pieces of the paths of all of the spoke ends. If the interstitial transformation was not diffeomorphic, as when objects slid along each other between individuals, an interpolation that allowed such transformations would need to be used.

We determine the fine scale warp to be composed with the transformation interpolated from the medial transformation using the fluid-flow warp method of (Miller et al., 1999). If the final map might not be diffeomorphic, as when regions of gas formed or were lost in the rectum or when tumors existed in the particular patient but the model was of well patients, then a warp method that permitted such situations would need to be applied.

The approach of computing a small scale space warp to be composed with a medially determined warp has the following advantages over computing the whole deformation as a space warp from an atlas. Optimizing large scale deformations is likely to be heavily affected by local minima, and in any case it is very likely to be slow as result of having to work over the combinatorially related, many small voxels. Indeed methods that have attempted to compute such warps have found it necessary to begin the process by preceding the voxel-scale warp by applying larger scale transformations such as ones based on manually chosen landmarks (Christensen et al., 1997). Using medial transformations to provide the large scale warp has advantages of being automatic, of using object-based correspondences, and of dividing itself into multiple scales, e.g., global to the object complex, object by object (with sympathetic inter-object relations), figure by figure (with sympathetic inter-figure relations), and medial atom by medial atom (reflecting inter-atom relations). Using these many scale levels produces both a much improved likelihood of convergence to the global optimum and qualitatively improved speed.

9.5.2 Evaluation of Segmentations

We have applied Pablo anecdotally to the segmentation of variety of organs or organ complexes. M-rep models have been built for both the liver (Han et al., 2005b), a multifigure object, and the heart (Pilgram et al., 2003), a multi-object ensemble, and statistical description of these anatomic entities have been generated. Controlled evaluations, described in the next two sections, have been carried out for the following two situations: (1) the extraction of kidneys from new patients' CT scans; (2) the extraction of the bladder, prostate, rectum complex from CT scans of a patient on one day of radiation treatment given the CT scans and segmentations of that patient on the planning day and other days of treatment. The first of these involves the segmentation of a single single-figure object with statistics drawn from many other patients' images, so we refer to it as an *inter-patient segmentation*. The second involves the segmentation of a multi-object complex with the statistics describing the variation across days within a patient (*intra-patient*).

9.5.2.1 Inter-Patient Kidney Segmentation Results

We have studied segmentation of the kidney from CT scans over a few years. An early result of evaluation of an earlier form of Pablo was described in (Rao et al., 2005). In that study we determined that averaged over a particular test sample, two humans' manual segmentations differed from each other in average surface distance over the kidney by 1.2 mm. Averaged over these cases the Hausdorff distance between the two segmented kidneys was 10 mm.

In a controlled study on segmentation of kidneys from 3D CT images of clinical quality, we used the sum of Mahalanobis distances described in Sections 9.3 and 9.4.2 as the objective function at the figure (object) stage. Since the log probability densities relieves the necessity of setting the relative weights of the two terms of the objective function by user control, these weights were set to unity in the study. However, at the atom stage we used the average squared-distance between each atom and the geodesic average of its neighbors, i.e, the atom irregularity penalty used in Binary Pablo (Section 9.3.1) for the geometric typicality, since the probability density training for the atom stage was not yet ready. This required a manually set weight on this term, which was held fixed for the experiment. The DRIQFs used in the geometry-to-image match at both stages were from Gaussian weighted regions inside and outside the kidney that had $\sigma = 3$ mm. In one trial training was on 20 cases and testing was done on 19 cases. In another trial leave-one-out testing was applied, i.e., all tests with 38 training cases and 1 test case were evaluated. In the geometry-to-image match, principal components carrying 97% and 99% of the variance were used to form the inside-object and outside-object log probability densities, respectively.

For our evaluation, we first consider the segmentation result to be that leading to minimum values of the atom-stage objective function. On the 19 test CTs the segmented kidneys had average surface distances to one human segmentation that

was at least as good as found between humans in the Rao study on a different test set. More precisely, the computer vs. human segmentations differed from each other in average surface distance over the kidney by 1.2 mm on the average case, and the Hausdorff distance between the two segmented kidneys was 6.8 mm on the average case. In all of the test cases, the automatic segmentation was usable without editing in radiotherapy treatment planning, although a voxel-stage editing would have been considered desirable in many of the cases. In fact, the automatic segmentations are frequently judged to be superior to the manual segmentations, and they have the additional benefit of being reproducible, even if the initialization is slightly different.

In the leave-one-out experiment, with its larger training sets, the results were roughly equivalent. The results of both experiments are given in more detail in (Broadhurst et al., 2006). An atom stage with a probabilistic geometric typicality might be expected to yield a further improvement.

These results were made less impressive by the fact that the objective function optimum that was found was not always achieved when we used the initialization based on six landmarks that we anticipated using for clinical purposes, namely, the north and south poles of the kidney and the two kidney crests at the level of the renal pelvis, and two positions at that level centered between the two crest landmarks. However, only 2 of the 19 cases in the first experiment and 4 of the 39 cases in the leave-one-out experiment would have required editing for clinical purposes.

9.5.2.2 Intra-Patient Multi-Object Male Pelvis Segmentation Results

As illustrated in Fig. 9.9, we have built a model for the multi-object ensemble of bladder, prostate, and rectum in the male pelvis. We have fit this multi-object model into a few dozen binary segmentations of these organs from fraction-by-fraction[2] CT images in five patients' cases, and after alignment of the prostate based on the two landmarks of the urethral entrance into and exit from the prostate, and after alignment of the bladder based on two polar and four equatorial landmarks, we have built statistical descriptions of the variability of the these objects across fractions within each particular patient. As well, we have built DRIQF statistics as described in the previous section, but here for six regions: interior and exterior regions for each organ. For the prostate and for the bladder, we also evaluated the use of approximately 200 overlapping regions to produce the exterior DRIQFs.

Finally, we have used these statistics to segment the prostate and the bladder from the CT images in other fractions in a leave-one-out experiment. (The rectum was represented as a tubular m-rep, and successful segmentation of the rectum was done in a separate experiment.) The initialization was done using the aforementioned landmarks. Since we are nearly ready to apply our method of principal geodesic analysis on medial atom residues and factor analysis on DRIQFS in local atom-relative object regions, we have optimized at the object stage only. The best results are produced when using the 200 exterior regions for DRIQFs. These results show

[2] A fraction is the radiation treatment on a given day.

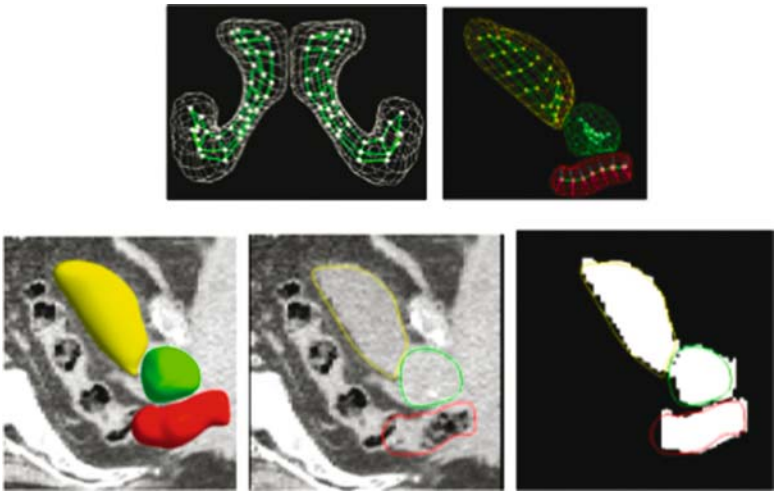


Fig. 9.9 M-reps for segmenting the male pelvis in CT images in later radiotherapy fractions. *Top left*: m-rep for pubic bones, used to register the later day fraction images with the first day fraction. *Top-right*: the m-rep for the bladder, prostate and rectum. *Bottom left*: a visualization of the bladder, prostate rectum m-rep's implied boundaries relative to a slice of the associated 3D CT image. *Bottom middle and right*: the segmentation result in a later fraction, shown in one of the image slices first vs. the greyscale CT image and then vs. the human segmentation shown in white

segmentations that have a median, over the cases, of the intersection/average volume overlap to a human segmentation of 93% for both the bladder and prostate and a median, over the cases, of the average closest point distance to the human segmentation of 1.13 mm for the bladder and 0.99 mm for the prostate. The numbers for the prostate, comparing segmentation based on statistics from a human who produced the training manual segmentations to the that human's result in the left-out-case, should be compared to the numbers comparing another observer's manual segmentation of the prostate to that of the training observer in one of the five patients' set of 16 multi-day CTs (Foskey et al., 2005). The agreement of the two humans' segmentations was 81% volume overlap and 1.9 mm average closest point surface separation.

When our segmentation was not as good as we wished, there were two explanations. First, in many of the segmentations of the bladder, a smaller scale refinement was necessary. We expect this to be accomplished when the log posterior optimizing atom stage is applied. Second, in a few cases the bladder initialization based on prostate landmarks was not adequate, but with a bladder-based initialization the segmentation was improved in a majority of cases.

This multi-object segmentation has been adapted for the clinical situation of adaptive radiotherapy by training the object principal geodesics by a pooling of aligned deviations from the mean of other patients. The results, which will soon be published with the details of the method, are comparable to those reported above.

Also, we expect shortly to report results of the atom-stage refinements of these segmentations.

Moreover, we are presently investigating having each object's change at the object scale level be divided into an m-rep change $\Delta\mathbf{m}^{self}$ independent of neighboring objects and an m-rep change $\Delta\mathbf{m}^{ngbr}$ reflecting the effect on the object of changes in neighbor atoms (Jeong et al., 2006). The neighbor-induced change is statistically described using the method of augmented object descriptions and prediction described in Chapter 8, Section 8. $\Delta\mathbf{m}^{ngbr}$ is decomposed as a conditional mean of the object, given designated neighbor atoms in its neighbors, plus a neighbor-effect residue with its own probability density. Probability densities on $\Delta\mathbf{m}^{self}$, on the augmented object, and on the neighbor effect residue are estimated by repetition of successive principal geodesic analyses. In segmentation the posterior is successively optimized with the prior iteratively in succession being the *self* probability density and the *neighbor residue* probability density, respectively. Initial results from statistical analysis on the bladder, prostate, rectum object complex are biologically reasonable, but it remains to test this approach by segmentations that use the self and neighbor residue probability densities.

9.5.2.3 Speed of Computation

The speed of a 3D segmentation on a Pentium IV, 1.7 GHz computer subdivides as follows.

- Preprocessing computations take less than 1 second.
- The largest scale stage (the object complex stage for an ensemble, the object stage for a single multifigure object, the figure stage for a single-figure object) takes a about 5 seconds per iteration and on the average requiring 20 iterations for a total time of about 3 minutes to determine the geometric warp coefficients.
- When the smaller scale medial stages are appropriately re-programmed, the same numbers will apply to each subfigure stage and about double for a full pass through the atoms at the atom stage, modulo the number of iterations required.
- The voxel displacement stage has not been timed, but it is expected to operate in under a minute.

Thus the total time for a kidney segmentation will typically be 7 minutes to segment a single-figure object.

While the method's speed has already benefited strongly from moving much of the computation from the deformation stage to the model building stage, there is still much room for speedup of integer multiples by more medial levels of coarse to fine, by medial deformation measured directly from the atoms without resort to the implied boundary, by having the gradients of the objective function relative to the changing parameters needed by the optimization steps be computed analytically rather than with numerical derivatives (shown in initial tests to more than double the speed), and just by more careful coding.

9.6 Hypothesis Testing for Localized Shape Differences Between Groups

We now focus on the quantitative morphologic assessment of structures between groups of human subjects. Our examples are individual brain structures in neuroimaging. Conventional methods study only volumetric changes, which explain intuitively global atrophy or dilation of structures. On the other hand, structural changes at specific locations are not sufficiently reflected in volume measurements. Statistical shape difference testing has thus become of increasing interest. Its potential to precisely locate morphological changes and to discriminate between groups makes it a good choice for studying pathological morphologic processes due to disease, as well as neuro-developmental processes. For example, we may wish to understand the shape differences in the hippocampus, caudate nucleus, cerebral ventricle complex in the brain between control patients and schizophrenics, or we may be interested in the differences of the hippocampus between 2-year olds who will develop autism and 4-year olds who will develop autism.

We focus in this section on the *hypothesis testing* of *whether* and *where* there are m-rep shape differences between the groups. We will discuss both global tests and truly local tests. Hypothesis testing applications using other medial descriptions have been proposed by Golland et al. (1999) and Bouix et al. (2005a).

We call the group designator C, which is numbered from 1 to the number of studied groups. Each group C_i consists of the objects of a sample of N_i cases. We assume that the objects or object complexes have been aligned across all cases, with the same alignment applied for the cases in both classes. The discrete m-rep objects are described as a tuple of medial atoms. The first idea is either to take all the atoms together and do a global test by studying the multivariate tuple of atoms $\times$ the 8 or 9 parameters per atom. Such a test can be powerful but will fail to localize the differences found to a particular collection of locations (i.e., parameters).

The alternative is to do a local (for a particular parameter of a particular atom) test on each atom parameter, at each position. We will use the term *location* to refer to such a combination of parameter and atom. The first idea might be to design a statistical test separately on any such location, and then to repeat that test over all atoms $\times$ parameters. However, the atoms are all correlated, and the parameter values are all correlated. To avoid unintended looseness in the threshold for rejecting the null hypothesis for any parameter, the threshold for rejection has to be adjusted for each parameter in a way reflecting the correlations. In Section 9.6.1.2 we describe a non-parametric permutation method to deal with this problem for individual parameters.

Section 9.6.2 will then focus on testing the full m-rep atom parameters jointly in symmetric space at a fixed scale. Finally, Section 9.6.2.2 will discuss why even atom by atom testing is not adequately local to the regions determined by the atoms and how to more appropriately handle locality.

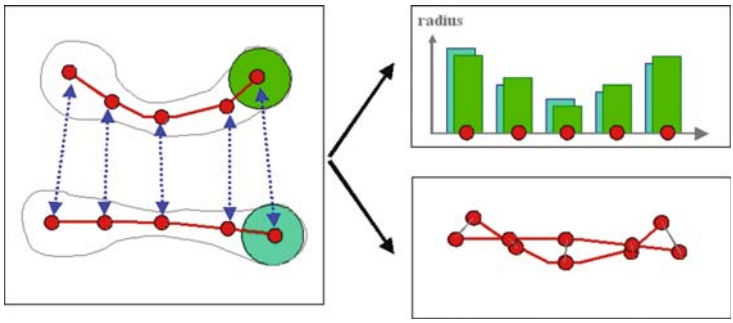


Fig. 9.10 Scalar m-rep shape difference (*schematically in 2D*) of 2 m-rep objects: Differences in the radius (*top graph*) and position (*lower graph*) are studied separately in Euclidean space. The properties express different kinds of underlying processes (*growth vs. deformation*) that are assumed to be statistically independent in the scalar testing

9.6.1 Tests in Euclidean Space

9.6.1.1 Univariate Tests in Euclidean Space

We may test a particular location (see Fig. 9.10). Here we focus on the two parameters, local position and radius (Styner et al., 2003b, 2004) of a particular atom. We first compute the group average objects by averaging the position and radius for each medial atom across each group. The overall average location is then computed as the average over all group average locations and serves as a template for computing univariate shape distance measurements. The signed position and thickness differences to the template are computed separately for the specified atom. The sign of the position difference is computed using the direction of the template medial surface normals.

Global shape analysis is computed by analyzing summarizing features such as the mean, median or other quantile measurements of the local differences across each object by standard statistical hypothesis tests. The choice of the feature evidently influences the outcome of the tests. The statistical tests mainly include parametric mean difference tests based on the Students-*t* distribution, and non-parametric mean difference tests, as well as parametric analysis of variance tests (ANOVA).

Local shape analysis does not need a summarizing feature as it is a truly local test. It is computed by testing each medial atom independently with a standard statistical hypothesis test. This results in a significance value (*P*-value) for each parameter and medial atom. We can represent this significance in a 3D visualization using color and size of spheres at the atom positions of the overall average object. This visualization, called a medial statistical significance map, allows one to locate significant shape differences between the groups in an intuitive but not truly local fashion (see Section 9.6.2.2). However, this raw significance map is incorrectly optimistic in regard to false-positive error rate because the atoms as well as the individual parameter values of a single atom are correlated, leading to the *multiple comparison problem*, a topic

of active research in the field of shape analysis (Worsley et al., 1996; Nichols and Holmes, 2001).

The raw significance map can be corrected for this multiple comparison problem using a uniformly sensitive, non-parametric permutation test approach (Pantazis et al., 2004) described in the next section. This results in a corrected significance map. In contrast to the raw significance map, which is a quite optimistic estimate of the real significance map, the corrected significance map is a somewhat pessimistic estimate, as discussed in the next section.

9.6.1.2 Multivariate Permutation Tests in Euclidean Spaces

The permutation tests we describe here localize regions (atom indices or parameters thereof) in objects that exhibit statistically significant morphological variation among two population groups while controlling the risk of any false positives, as long as the object features are in a Euclidean space. We find local thresholds that control the false-positive error rate and at the same time achieve uniform sensitivity among all locations.

We assume we have two groups of local parameter sets, group A and group B. Each parameter set represents either shape parameters or differences of shape parameters. We want to test the two groups for difference in the means at each location. Permutation tests are a valid and tractable approach for such an application. Our null hypothesis is that the distribution of the parameter set at each element is the same for every subject regardless of the group. Permutations among the two groups satisfy the exchangeability condition, i.e., they leave the distribution of the statistic of interest unaltered under the null hypothesis. Given n_1 members of the first group a_k, $k = 1 \ldots n_1$ and n_2 members of the second group b_k, $k = 1 \ldots n_2$, we can create $M \leq \binom{n_1+n_2}{n_2}$ permutation samples. A value of M from 20,000 and up should yield results that are negligibly different from using all permutations (Edgington, 1995).

The complex set of steps needed to test the null hypothesis that the two groups have the same probability distributions is illustrated in Fig. 9.11. We take the reader through this process step by step. For each permutation sample j, we compute a difference metric $\boldsymbol{T}_j$ between the groups, with elements T_{ij}. For univariate Euclidean parameters the absolute distance between the means of the groups is often used:

$$T_{ij} = \left|\hat{\mu}_{aij} - \hat{\mu}_{bij}\right| \tag{9.1}$$

where i is the location index, j the permutation index. If we wish to sense locations at which differences of collections of parameters at the locations are signficant, we can use difference metrics for multivariate, Euclidean or non-Euclidean parameters, as long as the difference metric itself is in Euclidean space, such as the multivariate Hotelling T^2 test statistic for the collection: $T^2 \propto D^2 = (\hat{\mu}_a - \hat{\mu}_b)^T \hat{\Sigma}^{-1} (\hat{\mu}_a - \hat{\mu}_b)$, where $\hat{\Sigma}$ is the pooled sample covariance. In $\mathbb{R}^n$ this statistic is invariant to coordinate transformations and is uniformly the most powerful test with this property

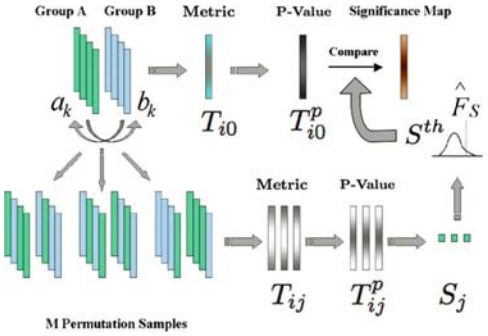


Fig. 9.11 Illustration of the permutation scheme. In the bottom row we create M permutation samples from the original data. We let j index the permutations and let i index the locations. For each permutation and location we compute the group difference metric T_{ij}, which is probability-normalized into T_{ij}^p. The data is then summarized across all locations to create the conservative summary statistic S_j over all locations. The empirical distribution of S_j, called $\hat{F}_S$ is used to define a global threshold S^{th} which for each location is applied to the probability-normalized test statistic obtained from the division to be tested, into groups A and B

(see Anderson, 1958 for a derivation). We cannot use this statistic directly on the multivariate combination of all atoms and parameters due to its inability to provide sensing of location.

In Fig. 9.11 it is assumed that we are given a target division of the data into two groups, A and B. To achieve uniform sensitivity across all locations, the parameter (or group of parameters) value T_{ij} at each location is first transformed to a uniformly distributed probability density value on [0,1], making all locations comparable. This is applied both to the test grouping, producing T_{i0}^p, and as illustrated in the bottom row of Fig. 9.11, it is also applied to the random permutations derived from the union of groups A and B, producing the T_{ij}^p. We can compare T_{ij}^p for each parameter i within each permutation j to produce a conservative summary statistic S_j for each permutation. Across the permutations the distribution of this summary statistic produces a common threshold S^{th} for each of the respective probability-normalized local parameters T_{i0}^p, as illustrated in the top row. The justification and specification of this scheme now follows.

The conservative summary statistic that we use for each permutation is the smallest probability density value over all locations i. We may then use the empirical distribution of this conservative summary statistic to extract thresholds that control the false positives to a desired level.

This method depends on having a form of normalization in the statistic T_{ij} that makes the locations comparable. A suitable normalization scheme is based on computing p-values, i.e., at each spatial location we compute the empirical distribution across permutations and then replace the statistic T_{ij} for each permutation sample with its p-value T_{ij}^p. The normalized metric T_{ij}^p is then guaranteed to have a uniform distribution on [0,1] under H_o for each i.

We can use the normalized data to define a local threshold map that controls the false-positive error rate to a desired level, say $\alpha = 5\%$, when applied to the original data. If the conservative summary statistic of the local parameters is $S_j = min_i\{T_{ij}^p\}$ over all locations i and $\hat{F}_S$ is the empirical cumulative distribution function of S, the appropriate global thresholds for a level α test would be $\hat{F}_S^{-1}(\alpha)$. For example, if we choose a threshold that leaves 5% of the area of the empirical distribution on the left side of S_j, then we have 5% probability of one or more false positives across all locations. This threshold S^{th} can be directly applied to T_{i0}^p (the statistic formed by probability-normalizing the original data with permutation index $j = 0$). Since the statistic T_{ij} is normalized separately for each location i, the same S^{th} corresponds to different values of local thresholds $p_i^{-1}(S^{th})$ of the unnormalized statistic T_{i0} at different locations.

Due to the use of the minimum p-value statistic across the whole surface, this correction scheme is focused only on controlling the rate of false positives at the given level α (commonly $\alpha = 0.05$) across the surface. No similar control of the rate of false negatives is available with this scheme. As the local significance level correctly controlled for false negatives can be anywhere between the raw p-value and the p-value corrected with our scheme, this corrected significance map is a pessimistic estimate of the true significance map.

9.6.2 Tests in Symmetric Spaces

The ideas in the previous section must be generalized to the non-Euclidean feature spaces appearing in m-reps and their symmetric space. We can derive permutation tests for equality of means of two groups using elements of the symmetric space. The sample means of each group, $\hat{\mu}_a$ and $\hat{\mu}_b$, can be computed using the techniques described in Chapter 8. Replacing T_{ij} from (9.1) with

$$T_{ij} = d(\hat{\mu}_a^*, \hat{\mu}_b^*) \tag{9.2}$$

yields a natural extension of local tests to symmetric spaces.

This provides a way to produce tests for a single aspect of the m-reps, such as position or radius of a particular atom, independently of the others, but typically we require a multivariate test using all of the parameters of one or more atoms simultaneously. We cannot fall back on Hotelling's T^2 test because it applies only to the linear case. Instead we can apply a transformation that forms new features from marginal probabilities, handling differing degrees of variability or correlation and making the test independent of magnification.

9.6.2.1 Global Multivariate Permutation Tests in Symmetric Spaces

We must now generalize the desirable properties of Hotelling's T^2 test to a nonparametric, nonlinear setting. One seemingly attractive option is to perform statistics

in the tangent plane as is done with principal geodesic analysis, since its linearity means Hotelling's T^2 test can be applied directly. However, with two samples, the question that arises is *which* tangent plane, since there is a different one around each sample's mean, and there may be no unique map between them.

The other conceptual problem is that if one follows geodesics past the *cut locus*—the set of points where two or more geodesics cross—then points on the manifold no longer have a single well-defined representative in the tangent plane. Instead of addressing these problems, we take a more general approach, which only requires that our objects lie in a metric space.

Our approach is based upon a general framework for nonparametric combination introduced by Pesarin (2001). The general idea is to perform a set of partial tests, each on a different aspect of the data, and then combine them into a single summary statistic, taking into account the dependence between the variables and the true multivariate nature of the data. When performing the partial tests, we require that each distribution has the same structure around the mean—equivalent to the assumption of a common covariance required by Hotelling—and test for a difference of means. More precisely, following the idea described in the previous section, we map each feature to its marginal probability and use these probability values as features. The following two sections describe the details.

Partial Tests. We compute test statistics T_{ij} as before, where as before i indexes the model parameters and j is the permutation index. We now turn to the case where we have Q test statistics: one for each of the parameters in our shape model. Let $\mu_{a,i}$ and $\mu_{b,i}$ be the means of the ith model parameter for each group. Then we wish to test whether any hypothesis $H_{1,i} : \{\mu_{a,i} \neq \mu_{b,i}\}$ is true against the alternative, that each null hypothesis $H_{0,i} : \{\mu_{a,i} = \mu_{b,i}\}$ is true. The partial test statistics $T_{ij}, i \in 1 \ldots Q, j \in 1 \ldots M$ are defined analogously to (9.2).

It can be shown that each of our mapped features T_{ij}^p has the properties of being significant for large values, consistent, and marginally unbiased, as defined in (Pesarin, 2001). Given that, Pesarin shows that a suitable combining function (described in the next section) will produce an unbiased test for the global hypothesis H_0 against H_1.

Since each of our tests are restricted to the data from a single model parameter and we have assumed that the distributions around the means in each group are identical, they are marginally unbiased. We cannot add an explicit test for equality of the distributions about the mean, as then the test for equality of means would be biased on its outcome.

To illustrate these ideas, we present a simple example, which we will follow through the next few sections. We take two samples of $n_1 = n_2 = 10$ data points from the two-dimensional space $\mathbb{R} \times \mathbb{R}^+$, corresponding to a position and a scale parameter. The samples are taken from a multivariate normal distribution by exponentiating the second coordinate, and then scaling both coordinates by a factor of ten. They are plotted together in Fig. 9.12a. They have a common covariance (before the exponentiation), and the two means are slightly offset in the second coordinate.

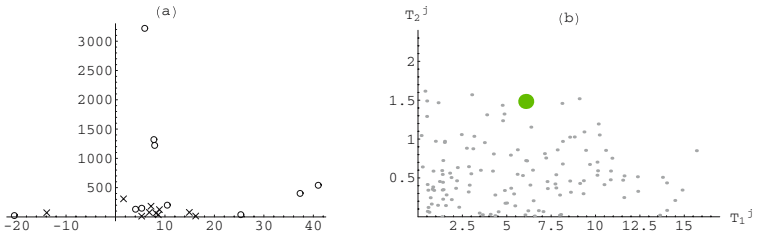


Fig. 9.12 The observed data and test statistics for our simple example. (**a**) shows the distribution of our two samples, with ×'s for the first and ○'s for the second. (**b**) shows the distribution of the partial test statistics under permutation. The large dot indicates the location of the observed data point

We construct $Q = 2$ partial test statistics using (9.2) for each coordinate, and evaluate them using Monte Carlo simulation with $M = 10{,}000$ permutations.

The results are shown in Fig. 9.12b. The first partial test value lies in the middle of the distribution, while the second lies near the edge. However, the scale of the first test is much larger, because no logarithm is involved in its metric.

Multivariate Combination. Given the partial tests from the previous section, we wish to combine them into a single test, while preserving the underlying dependence relations between the tests. This is done in the following manner. We apply the same M permutations to the data when computing each of partial tests, and we then compute a p-value statistic, T_{ij}^p as in Section 9.6.1.2. It is critical to use the same permutations for each partial test, as this is what captures the nature of the joint distribution.

We now wish to design a combining function to produce a single summary statistic, T_j', from each p-value vector $\boldsymbol{T}_j^p$. For one-sided tests, this statistic must be monotonically non-increasing in each argument, must obtain its (possibly infinite) supremum when any p-value is zero, and the critical value S'^{th} must be finite and strictly smaller than the supremum. If these conditions are satisfied along with those on the partial tests from the previous section, T_j' will be an unbiased test for the global hypothesis H_0 against H_1 (Pesarin, 2001).

Our combining function is motivated by the two-sided case (with signed distances), where we can use the Mahalanobis distance. Thus we need to transform the uniformly distributed p-values to a random variable that is normally distributed with mean zero and standard deviation 1. This is straightforwardly accomplished by applying the inverse of the cumulative density function for that Gaussian after subtracting $\frac{1}{2M}$. The extra $\frac{1}{2M}$ term keeps the values finite when the p-value is 1, and it is negligible as M goes to infinity. That is, we compute a $\boldsymbol{U}_j$ vector for each permutation, where $U_{ij} = \Phi^{-1}(T_{ij}^p - \frac{1}{2M})$, $j \in 1 \ldots M$, and Φ is the cumulative distribution function for the standard normal distribution. The map via the p-values and the Φ function gives the statistics known distributions that are directly comparable.

Arranging the $\boldsymbol{U}_j$ vectors into a single $M \times p$ matrix $\mathbf{U}$, we can estimate the covariance matrix $\hat{\Sigma}_U = \frac{1}{M}\mathbf{U}^T\mathbf{U}$ and use the Mahalanobis statistic: $T_j' = (\boldsymbol{U}_j)^T \hat{\Sigma}_U^{-1} \boldsymbol{U}_j$.

In the event that the data really is linear and normally distributed, the $\hat{\Sigma}_U$ matrix converges to the true covariance as the sample size goes to infinity (Pallini and Pesarin, 1992), making it asymptotically equivalent to Hotelling's T^2 test. Even if the sample size is small, the matrix Σ_U is well-conditioned regardless of the number of variables, since it is the covariance over the M permutations.

Typically, our distances are not signed, so we are interested in a one-sided test. In this case, we use the positive half of the standard normal cumulative distance function, $U_{ij} = \Phi^{-1}(1 - \frac{1}{2}(T_{ij}^p - \frac{1}{2M}))$ and assume the $\boldsymbol{U}_j$ distribution is symmetric about the origin. This assumption, however, implies that the covariance between $U_{i_1 j}$ and $U_{i_2 j}$ is exactly zero when $i_1 \neq i_2$. The diagonal entries of $\hat{\Sigma}_U$ are 1 by construction, so $\hat{\Sigma}_U = I$, the identity matrix. The fact that the p-values of the partial tests are invariant to scale obviates the need for arbitrary scaling factors. Thus, our one-sided combining function is

$$T'_j = (\boldsymbol{U}_j)^T \cdot \boldsymbol{U}_j. \tag{9.3}$$

The normality of the partial test statistics is not required. Also, even though the marginal distributions of the $\boldsymbol{U}_j$ vectors are normal, the joint distribution may not be. Therefore, we must use the empirical distribution of T'_j in order to compute the final p-value of the global test: $T_0'^p$. It is this nonparametric approach that corrects for correlation among the tests, even without explicit diagonal entries in the covariance matrix.

We return to our example from the previous section. The $\boldsymbol{U}_j$ vectors are plotted in Fig. 9.13a, along with the $\alpha = 0.05$ decision boundary, and our sample is shown to lie outside of it. As can be seen, equal power is assigned to alternatives lying at the same distance from the origin in this space. Figure 9.13b shows this boundary mapped back into the space of the original p-values.

The entire procedure is very similar to procedures used in correction for multiple tests described in the previous sections. However, instead of trying to find a local

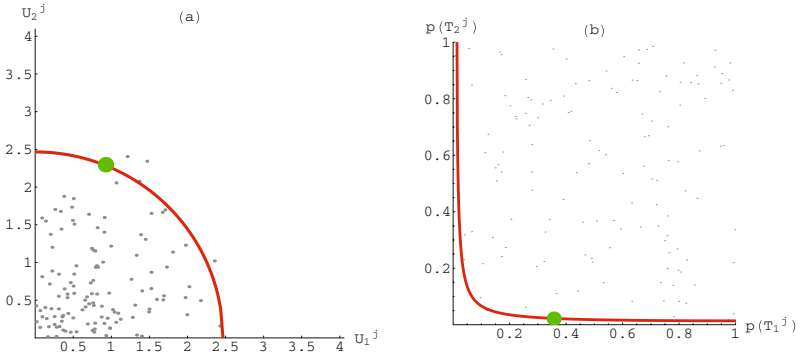


Fig. 9.13 The empirical distribution of our example plotted against the decision boundary at $\alpha = 0.05$. (**a**) The distribution of the $\boldsymbol{U}_j$ vectors, where the cutoff is a circle centered around the origin. (**b**) The distribution of the original p-values with the decision boundary pulled back into this space. The observed value is shown as the large dot in both plots

threshold for each test individually, we carve out a region of the multivariate T_{ij}^p space that contains some particular fraction, e.g., 5%, of the data to label as significant. We lose the ability to say *which* test is significant but gain power in the cases where multiple statistics independently signal significant differences.

9.6.2.2 Local Multivariate Tests in Symmetric Spaces

A test on all of the geometric primitives (e.g., medial atom) taken together is not truly a large scale test, for it confuses correlation with spatial scale. A test on each geometric primitive is not truly a small scale test, for it will respond equally well to a variation with large spatial scale as to one with a small scale. The Markov assumption on geometric neighbors allows the separation of scales by removing the correlation of neighboring elements from an element. In particular, if we can estimate the best predictor of a primitive by its neighbors and subtract that predictor from the primitive, the resulting residue provides the entity to test for significant variation *at the specified locality*.

This idea can be used for primitives such as objects or figures, but we are presently working to test it at the scale of the medial atom. Using the ideas in Section 9.1, the hypothesis testing would thus be done on each geodesic difference of the interpoland from the atom. However, we are still working on this form of local test, so the following section simply tests the atoms, one by one, rather than their residues.

9.7 Applications of Hypothesis Testing to Brain Structure Shape Differences in Neuro-Imaging

This section presents two application of hypothesis testing of m-rep objects. In the first application, scalar hypothesis testing of individual medial parameters was employed (see Section 9.6.1.2) for analyzing hippocampal shape in schizophrenia. In the second application, hypothesis testing in the symmetric space (see Section 9.6.2) was employed for analyzing ventricular shape in healthy twins and in schizophrenia.

9.7.1 Hippocampus Study in Schizophrenia

In the study presented in this section, we investigated the shape of the hippocampus structure in the left and right brain hemisphere in schizophrenic patients (SZ, 56 cases) and healthy controls (Cnt, 26 cases). The hippocampus is a gray matter structure in the limbic system and is involved in processes of motivation and emotions. It also has a central role in the formation of memory. Hippocampal atrophy

has been observed in studies of several neurological diseases, such as schizophrenia, epilepsy, and Alzheimer's disease. The goal of our study was to assess shape changes between schizophrenic patients and the control group.

The subjects in this study all have male gender and the same handedness. The two populations are matched for age and ethnicity. The hippocampi were segmented from inversion-recovery-prepped SPGR MRI datasets (resolution: $0.9375 \times 0.9375 \times 1.5$ mm) using a manual outlining procedure based on a strict protocol and well-accepted anatomical landmarks (Duvernoy, 1998). The segmentation was performed by a single clinical expert (Schobel et al., 2001) with intra-rater variability of the segmented volume measurements at 0.95. Spherical harmonic (SPHARM) coefficients were computed using a sampling of 2,252 points, and the results were normalized via a rigid-body Procrustes alignment and a scaling to unit volume. The m-rep model was built on the aligned full population including the objects of all subjects on both sides, with the right hippocampi mirrored at the interhemispheric plane prior to the model generation. The resulting m-rep model has a single figure topology and a grid sampling of 3×8 medial atoms, in total 24 atoms. The range of the average distance error between the fitted m-rep's boundary and the original boundary was between 0.14 and 0.27 mm (mean error 0.17 mm), less than half the voxel size of the original MRI, so the medial shape analysis should capture the shape changes in the image data.

The template for the medial shape analysis was determined by the overall average structure. As the two populations are not equal in size, we computed the overall average as the average of the population averages. Due to age variation in both populations, the shape difference values were corrected for age influence, using a linear least squares model.

The global shape analysis in Table 9.1 shows a strong trend in the m-rep position analysis on the left side. The m-rep thickness analysis is significant for neither hippocampus. This suggests a deformation shape change in the hippocampus between the schizophrenic and the control group. The results of the m-rep position analysis shows a stronger significance than the SPHARM-PDM analysis that was also carried out. Additionally to the mean difference, several quartile measures (Median, 75% and 95%) were analyzed and produced structurally the same results.

The m-rep local position shape analysis (Fig. 9.14) yields significant changes that are in roughly the same position, mainly in the hippocampal tail, as shown by SPHARM-PDM shape analysis and by distance maps of the averages. No significance was found in the local m-rep thickness analysis. Similar to the outcome of the global analysis, the local m-rep position analysis shows a stronger significance

Table 9.1 Results of global shape analysis (average across the surface/medial manifold): Table of group mean difference p-values between the schizophrenic and control group ($*$: significant at $\alpha = 0.05$ significance level)

Global analysis	M-rep thickness	M-rep position
Left	$p = 0.722$	$p = 0.0513$
Right	$p = 0.751$	$*p = 0.0001$

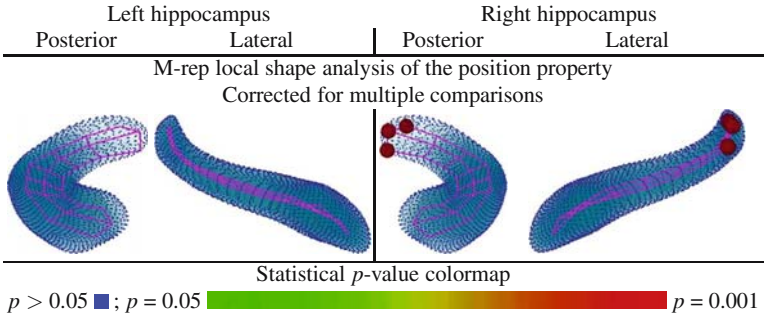


Fig. 9.14 Statistical maps of the local shape analysis from posterior and lateral views, corrected for multiple comparisons. The m-rep analysis shows the statistical significance at each medial atom using both the color and the radius of spheres placed at the atom positions. The main area of significance is located at the hippocampal tail. The corrected results are overly pessimistic

than the SPHARM-PDM analysis. The local shape differences are mainly located at the right hippocampal tail, with near significance in the left hippocampal tail. By inspecting the average structures of the two groups, we further find that the hippocampal tail region of the control group in our study is more bent than the one of the schizophrenic group.

9.7.2 Lateral Ventricle Study of Healthy and Schizophrenic Twins

The data for these experiments comes from a twin pair schizophrenia study conducted by Weinberger et al. (2001). High resolution $(0.9375 \times 0.9375 \times 1.5\,\mathrm{mm}^3)$ MRIs scans were acquired from three different subject groups: 9 healthy monozygotic twin pairs (MZ), 10 healthy dizygotic twin pairs (DZ), and 9 monozygotic twin pairs with one twin discordant for schizophrenia and one twin unaffected (DS). See Fig. 9.15 for some examples. A fourth group of 10 healthy non-related subject pairs (NR) was constructed by matching unrelated members of the two healthy groups. All four groups were matched for age, gender, and handedness. A tenth healthy, monozygotic twin pair was discarded due to possible brain shape changes attributed to major head trauma suffered by one of the twins in a car accident at age seven. A tenth twin pair discordant for schizophrenia was discarded due to hydrocephaly in the unaffected twin.

The left and right lateral ventricles were segmented using automatic atlas based tissue classification (van Leemput et al., 1999) and 3-D connectivity. An automatic morphological closing operation was applied to ensure a spherical topology. An area-preserving map was used to map them to a sphere, after which they were converted to a spherical harmonics representation (SPHARM) (Brechbühler et al., 1995). Correspondence on the boundary was established using the first order harmonics (Gerig et al., 2001). Point Distribution Models (PDMs) were constructed by

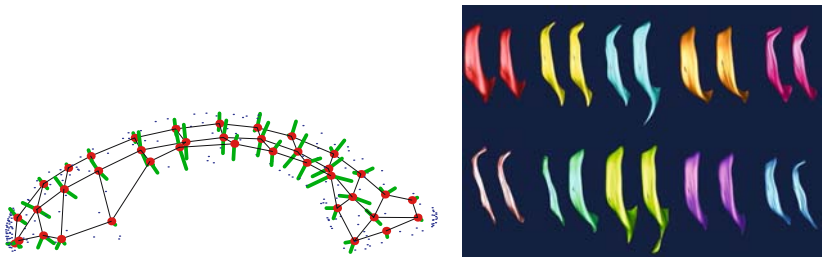


Fig. 9.15 *Left*: An example m-rep of a left lateral ventricle. The mesh vertices and off-shooting spokes make up the medial atoms. The shape the m-rep was fit to is shown as a point cloud surrounding it. *Right*: Ventricle pairs from five monozygotic twin pairs (*top*) and five dizygotic twin pairs (*bottom*)

uniformly sampling the boundary at corresponding points. The m-rep models were constructed using a robust method that ensures a common medial topology (Styner et al., 2003a). For our data, this consists of a single medial sheet with a 3×13 grid of medial atoms, which provides 98% volume overlap with the original segmentations.

From this data set, we wish to determine if the twin pairs that were more closely related had smaller variations in shape. We also wish to see if the shape variations between the discordant and the unaffected twins in the schizophrenic pairs is similar to the normal variation between healthy monozygotic twins. For this purpose, we use the partial test statistics:

$$T_{ij} = \frac{1}{n_1} \sum_{k=1}^{n_1} d(a_{ki}^{1*}, a_{ki}^{2*}) - \frac{1}{n_2} \sum_{k=1}^{n_2} d(b_{ki}^{1*}, b_{ki}^{2*}). \tag{9.4}$$

Here (a^1, a^2) form the twin pairs for one group, while (b^1, b^2) form the twin pairs for the other. The partial tests are applied separately to all three components of the medial atom location, $\mathbf{x}$, as well as the radius and two spoke directions. This gives six partial tests per medial atom, for a total of $p = 3 \times 13 \times 6 = 234$, much larger than the sample size. Each is a one-sided test that the variability in group 2 is larger than that in group 1.

For consistency with previous studies (Styner et al., 2005), all shapes were volume normalized. After normalization, we also applied m-rep alignment, as described by Fletcher et al. (2004), to minimize the sum of squared geodesic distances between models in a medial analog of Procrustes alignment. First, the members of each twin pair were aligned with each other, and then the pairs were aligned together as a group, applying the same transformation to each member of a single pair.

In order to ensure invariance to rotations, we had to choose data-dependent coordinate axes for the $\underline{\mathbf{x}}$ component of each medial atom. Our choice was the axes which diagonalized the sample covariance of the displacement vectors from one twin's atom position to the other at each site. While this had some influence on the results, the general trend was the same irrespective of the axes used.

Table 9.2 *p*-values for paired tests for the difference in the amount of shape variability in groups with different degrees of genetic similarity. Results from our method are in the first two columns, while results from a previous study (Styner et al., 2005) are in the last two for comparison. Groups are: monozygotic (MZ), monozygotic twins with one twin discordant for schizophrenia (DS), dizygotic (DZ), and non-related (NR) ($*$: significant at the $\alpha = 0.05$ significance level).

	Our study		Boundary study (Styner et al., 2005)	
	Left	Right	Left	Right
MZ vs. DS	0.12	0.38	0.28	0.68
MZ vs. DZ	$*$0.00006	$*$0.0033	$*$0.0082	$*$0.0399
MZ vs. NR	$*$0.00002	$*$0.00020	$*$0.0018	$*$0.0006
DS vs. DZ	$*$0.020	$*$0.0076	0.25	0.24
DS vs. NR	$*$0.0031	$*$0.00026	$*$0.018	$*$0.0026
DZ vs. NR	0.16	0.055	$*$0.05	$*$0.016

For each pair of twin groups, we generated $M = 50{,}000$ permutations, and computed their *p*-value vectors. Following Section 9.6.2.1, these were mapped into $\boldsymbol{U}_j$ vectors, from which the empirical distribution of the combined test statistic T'^k from (9.3) was estimated, producing a single global *p*-value.

The results are summarized in Table 9.2. For comparison, we list the results of a previous study which used a univariate test on the average distance between corresponding points on the PDMs (Styner et al., 2005). While we note that the significance of a *p*-value on an experimental data set is not a useful metric for comparing different methods, it is interesting to see the differences between the two. Our tests give a consistent ranking: MZ $\approx$ DS $<$ DZ $\approx$ NR, which is fully transitive. The boundary study, however, finds a significant difference between DZ and NR, but fails to identify the difference between DS and DZ.

We also performed local tests, to identify specific medial atoms with with strong differences. A multivariate test was conducted using our procedure on the 6 components of each atom, and the results were corrected for multiple tests using the minimum *p*-value distribution across the shape described in Section 9.6.1.2. The results are shown in Fig. 9.16.

9.8 Discussion and Future Work

9.8.1 Are M-Reps Effective?

The main objective of this chapter was to describe m-reps based methods for 3D medical image segmentation and for statistical characterization of differences of anatomic shapes seen in populations of medical images. M-reps have been used both to capture knowledge of object geometry and to give a basis of the positional correspondences needed in training and measuring geometry-to-image match log probabilities. As well, they have allowed efficient, multiscale operation in both training the probabilities and applying them. It has been our expectation that they provide

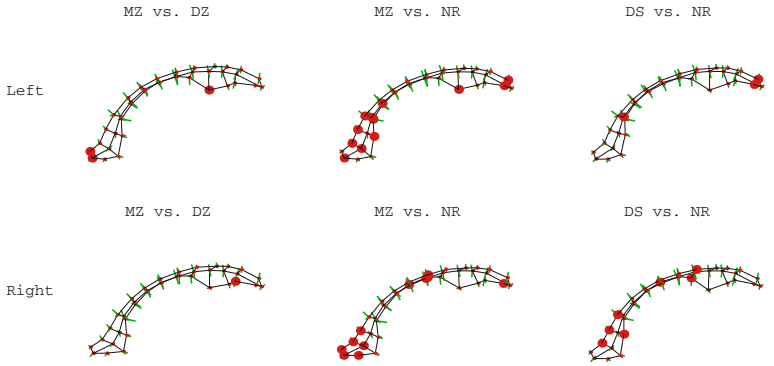


Fig. 9.16 Results for local tests for the difference in shape variability in groups with different degrees of genetic similarity. Atoms with differences significant at the $\alpha = 0.05$ level are shown in a larger size. Tests not shown had no significant local differences

more stable estimates of modes of variation and the associated principal variances for a given number of training samples than alternative bases for geometric statistics, and we have some early results suggesting that this is the case, but this is yet to be proven.

In addition, much more than other geometric representations, m-reps have provided a means of yielding probability distributions whose samples were very unlikely to be geometrically improper, avoiding illegal interpenetrations and creases. Checks on geometric propriety via S_{rad} has avoided creasing or near creasing and the improved estimates of boundary normals without lowering the tightness of boundary fits to training binary images. DRIQF statistics based on these fits have led to improved segmentations.

The success of m-reps as an object representation designed for statistical uses should be judged by the success of the applications. Within the class of deformable models methods that might be considered to provide comparable segmentation accuracy, robustness and low interaction requirements, m-reps based segmentations are among the speedier.

In terms of accuracy and robustness, the 3D segmentation method based on m-reps have produced single-figure object, viz. kidney, segmentations that are competitive with human manual segmentation and are, to our knowledge, the most accurate kidney segmentations reported in the literature. The same can be said of the initial multi-object segmentations of male pelvis objects in CT images using intra-patient statistics, though given the serious challenge of this problem, further work must be done before the method can be tested on many patients and its results compared to the results of alternative methods for segmentation of these objects. Moreover, while the apparatus for segmentation of multi-figure objects exists and has been tried on simple test cases, real application and testing of such segmentation is yet to be done.

Of course, when comparing m-reps to other object representations that are being used for segmentation via deformable models, the issue is not simply whether

m-reps are as good or better than the alternatives, but whether they are enough better to justify the complexities of the medial representation. Controlled, quantitated validation on a variety of objects by multiple methods in competition needs to be carried out before this can be judged.

We are in the process of making the following improvements to our deformable m-reps segmentation method and software:

1. Sensing and reporting locations on the segmented object that do not have the expected level of geometry-to-image match, so that the user can take actions of relocating that object section and then restart the segmentation.
2. Bringing to routine usability a posterior optimizing atom stage as well as the option of computing a small scale diffeomorphism both in the target object(s) and in the interstitial space between objects in place of the small scale boundary displacement.
3. Developing a form of our software intended for clinical use and thus being as automatic as possible, and making all interactions in clinical terms.

The m-rep hypothesis testing tools have been applied to several studies in neuroimaging and have shown to provide meaningful results. The main advantage of our m-rep hypothesis testing tools over boundary based testing tools is the identification of different types of processes using the different m-rep atom properties. This leads to results that are more intuitively interpretable. In several studies of the hippocampus, the caudate and the lateral ventricles, we have shown that the overall results correlate well between medial and boundary description, but also that our m-rep analysis is able to capture additional information not seen in the boundary analysis.

Our current hypothesis testing tools are based on a true multivariate permutation test approach for hypothesis testing in direct products of metric spaces. The resulting test does not require a priori scaling factors to be chosen, and captures the true multivariate nature of the data. It is well-defined even in the high-dimensional, low-sample size case. The method has been developed for m-reps, though it is suitable for any type of metric data, including potentially categorical data. An important area for future research is the design of suitable partial tests to use in each space. Because they cannot be broken into smaller pieces than a single component of the direct product, the distance to the mean and similar tests are limited in the types of distributions they can describe. For example, the distance from the mean can only characterize an isotropic distribution on the sphere. This would allow us to relax our assumption of identical distribution about the mean.

Even though our hypothesis testing tools have matured to a degree that they can be employed routinely in neuro-imaging studies, there are several limitations to our current tools making the following enhancements to our methods necessary:

1. Developing a combined analysis of multiple objects in order to capture correlated differences of the shape in neighboring brain structures such as the lateral ventricle and the caudate.
2. Enhancing the analysis scheme to incorporate several layers of scale starting at the global multi-object scale down to the local single atom scale.

3. Incorporating statistical models of patient covariates such as gender, age and medication in the permutation test algorithm. The current method incorporates covariates by correcting atom parameters independently using least-squares linear regression.

9.8.2 Other M-Rep Uses and Properties

In a separate paper (Crouch et al., 2003) we have shown how the space parametrization provided by m-reps also allows the interior of the object to be divided into mesh elements useful for efficient mechanical modeling of intra-patient motion of anatomic structures due to such interventions as intrarectal imaging probes. The measures of mechanical energy computed in this approach could be used for segmentation of a patient whose segmented m-rep from an earlier (e.g., planning) image can be used as the model for a segmentation in a later (e.g., intra-treatment) image.

M-reps provide one means of modeling objects and collections of objects; boundary representations (b-reps) are a common alternative means of such object modeling. They share the limitations of all object modeling methods, namely that a single object model will not serve for a class of objects with mixed topologies at the figural level. However, because they explicitly model the interfigural relations, they have special weaknesses when these relations are variable over the population of objects. For example, an m-rep for a right kidney and a separate left kidney will not perform well for a horse-shoe kidney, in which the kidneys are joined. For such mixed classes, a separate m-rep is required for each exemplar. Another issue shared with other object models is instability for nearly spherically or circularly symmetric objects. In such cases the nearly degenerate geometry creates computational instabilities in discriminating among the three major axes which in turn can cause an m-rep to "flip" during deformation in the image data in an unstable manner. However, m-reps share with other object models the particular strength of resolving these orientational instabilities via the relations among objects.

M-reps' special abilities relative to b-reps derive from their explicit representation of object orientation changes such as twisting and bending and of object size changes such as widening and narrowing. Thus statistics on rectal widenings due to gas, on the variability in the relative pose of the two lobes of the liver, and on the orientation of the bladder relative to the prostate are very effective in m-reps terms. The limitations not of m-reps by themselves but of m-reps with statistics come in situations when the orientational or magnificational relationships are very variable. Thus, like b-reps m-reps are well suited to complex slabs and tubes such as the cerebral cortex or the intestine, and both are well suited to intra-patient variations of these structures over time. But because in the population of humans the variability of the folding structure of the cerebral cortex is high and the variability of the curvature of the intestine is high, statistics on m-reps is a weak tool over that population for these structures.

Because m-reps represent the interior of objects, they lose their effectiveness in image situations where only one side of an object appears in an image, and they have weakness relative to b-reps in situations where one side of an object boundary is statistically stable but the other side has great variability. In that situation b-reps can ignore the unstable or unimaged side, whereas m-reps inherently must represent both sides together.

M-reps allow statistics by providing a fixed topology of sheets and their branching. As presently designed, populations that are not well modeled by fixed topology m-reps together with voxel scale refinements will require a different geometric representation.

Acknowledgments This work was done under the partial support of NIH grants P01 CA47982 and P01 EB02779 and a grant by the Stanley Foundation. A gift from Intel Corp. provided computers on which some of this research was carried out. We thank Xifeng Fang, Qiong Han, Ja Yeon Jeong, Joshua Levy, Conglin Lu, Derek Merck, Joshua Stough, Gregg Tracton, Guido Gerig, A. Graham Gash, and Delphine Bull for help with models, figures, software, and references. We are grateful to J. Stephen Marron, Keith Muller, and Surajit Ray for help with statistical methods, to Jeffrey Lieberman, Julia Fielding, and Valerie Jewells for medical images, and to members of the UNC Neuroimage Analysis Laboratory and Julian Rosenman for manually segmenting images.