

Comparison of Human and Automatic Segmentations of Kidneys from CT Images

Rao* M, MS, Stough[§] J, BS, Chi[†] Y-Y, MS, Muller[†] K, PhD,
Tracton* GS, BS, Pizer[§] SM, PhD and Chaney* EL, PhD

*Department of Radiation Oncology

[§] Department of Computer Science

[†]Department of Biostatistics

University of North Carolina

Chapel Hill, North Carolina 27599-7512

Research supported by NCI P01 EB002779

ABSTRACT

Purpose: A controlled observer study was conducted to compare a method for automatic image segmentation with conventional user-guided segmentation of right and left kidneys from planning CT images.

Materials and Methods: Deformable shape models called m-reps were used to automatically segment right and left kidneys from twelve target CT images and the results were compared to careful manual segmentations performed by two human experts. M-rep models were trained based on manual segmentations from a collection of images that did not include the targets. Segmentation using m-reps began with interactive initialization to position the kidney model over the target kidney in the image data. Fully automatic segmentation proceeded through two stages at successively smaller spatial scales. At the first stage a global similarity transformation of the kidney model was computed to position the model closer to the target kidney. The similarity transformation was followed by large scale deformations based on principal geodesic analysis (PGA). During the second stage each medial atom comprising the m-rep model was deformed in turn and iteratively. The transformations and deformations at both stages were driven by optimizing an objective function with two terms. One term computed the geometric typicality of the currently deformed m-rep via the probability distribution of transformations and deformations derived from PGA of the training segmentations. The second term computed a model-to-image match term based on the goodness of match of the trained intensity template for the currently deformed m-rep with the corresponding intensity data in the target image. Human and m-rep segmentations were compared using quantitative metrics provided in a toolset called VALMET. Metrics reported in this paper include 1) percent volume overlap; 2) mean surface distance between two segmentations; and 3) maximum surface separation (Hausdorff distance).

Results: Averaged over all kidneys the the mean surface separation was 0.12 cm, the mean Hausdorff distance was 0.99 cm, and the mean volume overlap for human segmentations was 88.8%. Between human and m-rep segmentations the mean surface separation was 0.18-0.19 cm, the mean Hausdorff distance was 1.14-1.25 cm, and the mean volume overlap was 82-83%,.

Conclusion: Overall in this study the best m-rep kidney segmentations were at least as good as careful manual slice-by-slice segmentations, and the worst performance was no worse than humans in the typical clinical setting. The mean surface separations for human-m-rep segmentations were slightly larger than for human-human segmentations but still in the sub-voxel range, and volume overlap and maximum surface separation were slightly better for human-human comparisons. These results were expected because of experimental factors that favored comparison of the human-human segmentations. In particular m-rep agreement with humans appears to have been limited largely by fundamental differences between manual slice-by-slice and true three-dimensional segmentation, imaging artifacts, image voxel dimensions, and the use of an m-rep model that produced a smooth surface across the renal pelvis.

Key words: treatment planning, image segmentation, kidney

Introduction

Three-dimensional radiation treatment planning (3D RTP) systems require a user-created model of the patient to localize and display objects of interest, position the isocenter(s) of the treatment beams, shape the radiation beams to conform to the outline of the target volume and avoid nearby sensitive tissues, incorporate tissue inhomogeneities into dose calculations, and compute volume-weighted metrics such as dose-volume histograms (DVHs) that are used for comparing competing treatment plans. The anatomical structures and tumor related objects comprising the patient model are defined by segmenting one or more volume images, usually CT and MR images. Due to the large number of departments practicing 3D RTP and the large number of patients undergoing 3D RTP every day, segmentation of medical images is a commonly performed clinical task that affects critical treatment decisions. It is likely that segmentation is performed more often as a clinical procedure in radiation oncology than for all the other medical specialties combined. Unfortunately current segmentation practice is inherently inefficient and expensive. Most methods in routine clinical practice are user-guided slice-by-slice contouring tools that require well trained users to achieve acceptable results for 3D RTP. Other flaws of current segmentation methods that tend toward sub-optimal treatment planning include intra- and inter-user variabilities [Dowsett 1992; Leunens 1993; Valley 1993; Kagawa 1997; Roach 1996; Algan 1995; Ketting 1997a, 1997b; Rasch 1999], the lack of practical approaches that fully consider all three spatial dimensions, and the inability to deal with ambiguous surface localization.

The development of automatic 3D segmentation methods is motivated by several considerations including economic pressure to improve efficiency and contain costs, and the clinical need to improve accuracy and reproducibility in order to steer user-directed planning decisions and inverse treatment planning algorithms consistently in the right direction. Deformable shape models is a general class that is showing great promise for automatic segmentation of normal anatomical structures. Kass [1987] first described a straightforward method based on deformable two-dimensional contours popularly known as snakes. A useful survey of snakes is found in the paper by McInerney and Terzopoulos [1996]. Collections of papers on early deformable models can be found in the book by ter Haar Romeny [1994], and proceedings of conferences such as CVRMed '95 [1995] and CVRMED-MRCAS [1997], Montagnat [1997], McInerney [1996a,b], Jones [1997], and Vehkomäki [1997]. However, in order for classic snake-like deformable contours to be robust and reproducible in the clinical setting, the initial guesses for shape and position of the target object essentially must be equivalent hand-drawn contours. This requirement effectively precludes the possibility of replacing hand contouring with snakes. Statistically grounded deformable shape models that can be trained to capture *a priori* information about the probability distributions of target object shapes overcome many problems presented by classic snake-like methods. The IEEE TMI [1999] special issue on model-based analysis of medical images has a collection of papers on a number of these methods.

The more sophisticated deformable shape methods use explicit geometric models to represent object shape. Such models represent *a priori* information that can be used in a statistical framework for matching the model against a target image. For objects with predictable shapes such as normal anatomical structures the model can be thought of as representing a shape that is typical for the target object. For example an m-rep is a model of the mean shape that can deform, within the limits imposed by the probability distribution on target shapes, to match the shape of a corresponding object in a target image. The statistical framework for driving the deformation is reviewed briefly below and discussed in greater detail by Pizer [20, 21], Fletcher [22], and Lu [23].

In this paper we discuss the results of an observer study comparing automatic and human segmentations of left and right kidneys from planning CT images. Kidneys were selected for this study because they are relatively unchallenging for trained humans to contour and thus an acceptable reference standard is easily defined, and because of their importance for treatment planning. They also are a challenging initial objective for automatic methods because they are located in a crowded soft-tissue environment with bony structures nearby.

Segmentation was performed in this study using medial models called m-reps [20, 21]. M-reps have a number of strengths that are well matched to the task of segmenting normal structures from medical images for radiotherapy treatment planning [24].

Materials and Methods

M-reps

Detailed discussions of the structure, building, training, and deformation of m-reps can be found in papers by Pizer [20, 21]. The first of these papers, published in *Int J Comp Vis*, has image analysis experts as its target audience and focuses on the m-rep representation and its deformations. The second of these papers, submitted for publication but available on the midag.cs.unc.edu website, focuses on users of deformable m-reps based segmentation and includes matters of the segmentation method and statistical aspects. For completeness and continuity brief discussions relevant to the kidney m-reps used in this study are presented below.

The simplest three-dimensional (3D) shape is a single figure without subfigures, i.e., indentations or protrusions. For this study the combined kidney parenchyma and renal pelvis were treated as a single figure m-rep. Such an object is described in a medial framework by a two-dimensional medial sheet that implies a three-dimensional surface (Figure 1). The medial sheet consists of a curved surface with two oppositely directed spokes at every point defining the object width. The two equal-length spokes extend above and below the sheet and touch points on opposite sides of the object surface. The surface implied by the medial sheet simply passes over the spoke ends. The edges of the sheet have in addition a sequence of crest spokes that generates the piece of object surface, forming a curving crest, which connects the surface passing over the above-surface spokes to the surface passing over the below-surface spokes.

For an m-rep model the medial sheet is represented by a grid of medial atoms as shown in Figure 1. The centers of the atoms, called hubs, lie on the medial sheet's surface. Atoms on the interior of the grid have two spokes of equal length that extend to and are perpendicular to patches on opposite sides of the implied surface. Each atom on the edge of the grid has a third spoke that bisects the other two spokes and defines the radius of curvature for the crest section of the implied surface corresponding to that atom.

For computational efficiency the number of rows and columns, and thus the number of atoms, in an m-rep is selected to be the fewest needed to capture the full range of shape variability over the population of shapes in the target population [25]. When the grid dimensions have been determined, the corresponding m-rep can be trained to represent a right or left kidney by training on a set of images that capture the full range of shape variability. Truth is defined in the training images by experienced humans who segment the kidneys using a slice-by-slice contouring method. Two types of training are necessary, geometric and intensity. Geometric training determines the mean shape of an object and the principal modes of shape variation using a method called principal geodesic analysis (PGA) [22]. PGA is performed in curved space to allow characterization of local twisting, bending, magnification, as well as displacement, of the object interior. PGA is similar to the familiar principal components analysis performed in flat Euclidean space. In general the m-rep model for a particular object is taken to be the mean m-rep determined by PGA.

The intensity training method used in this study examined the intensity variation at 2562 locations over the implied surface of the kidney [26]. The locations were defined to be the vertices of the tiles that form the kidney surface computed using a modified subdivision surface method [27, 28] applied to the grid atom spoke end positions and directions. The relative intensity variation at each location, called an intensity profile, was measured at eleven points along line segments that passed through the vertices (Figure 2). The line segments were half inside and half outside the kidney and orthogonal to the kidney surface. The actual intensity profile at a particular vertex was measured across all training images and the resulting collection was compared to three canonical forms most representative of the profiles in the training data. The three forms were 1) light to dark,

capturing kidney boundary locations abutting darker fat, etc.; 2) dark to light, capturing kidney boundary locations abutting lighter liver, bone, etc.; and 3) a notch, capturing kidney boundary locations with a small amount of darker fat, etc. between the kidney and another section of organ tissue or bone. For this study the characteristic profile identified with a particular vertex was defined to be the most popular profile at that vertex over all training cases.

When a deformable model is placed in a target image it changes shape to match the corresponding object. Deformation is performed at multiple spatial scales. At the largest scale an m-rep model is translated and rotated as a whole to best match the location and pose of the target object. This step is followed by global surface deformations that change the whole grid of atoms to better match the target. The large scale deformations are computed from the principal modes of variation determined at geometric training. Large scale deformations are followed by atom-by-atom deformations. The deformation of each atom affects only a moderately broad interior object section surrounding the atom. A final boundary stage displaces individual surface points to achieve a fine-scale match with the target. The boundary stage captures fine detail and is best suited for "clean" images where the edge of the target object is well imaged and free of artifacts. In this study the target images contained significant imaging artifacts that could result in irregular surfaces at the boundary stage. To avoid capturing these artifacts the boundary stage was omitted, a decision that introduced bias favoring human-human comparisons because, as discussed later, human segmentations tend to preserve the artifacts present in the target images used in this study. All of the preceding steps occur in ~2-3 minutes per kidney on a 2003 vintage laptop.

Each stage in the m-rep deformation process is driven by optimizing an objective function that is the sum of two terms. The geometric typicality term measures the goodness of match between the current deformed state of the m-rep and the mean m-rep. This geometric term penalizes the current shape in proportion to its deviation from the mean. The image match term measures how well the intensity pattern in the target image data matches the intensity pattern of the characteristic profiles associated with the m-rep model.

Artifacts, discussed below, in the target images caused the renal pelvis to be poorly imaged. To produce consistent segmentation results in this region a single figure m-rep with a 5×3 grid of atoms was selected for use in this study (Figure 1) and the model was trained to produce a relatively smooth segmentation across the renal pelvis. Had our objective been to exclude the renal pelvis in the final segmentation we would have selected a multi-figure model comprising a main figure for the kidney and a subfigure indentation for the pelvis. The decision to include part of the renal pelvis in m-rep segmentation further biased human-human comparison because the human contours often demonstrated indentations at the renal pelvis.

Target and training images

The target images were a set of 12 planning CT images (24 kidneys in all) obtained from local department archives. The scans were collected using a Siemens Somatom Plus 4 CT scanner. The image matrix was 512×512 , the slice thickness was 5 mm, and the pixel size ranged from $0.098 \text{ mm} \times 0.098 \text{ mm}$ to $0.156 \text{ mm} \times 0.156 \text{ mm}$. The primary criteria for image selection were both kidneys had to be completely imaged with 2 cm superior and inferior margins, no contrast media, and slice thickness ≤ 5 mm. The protocol for acquiring the planning CT images used in this study involved non-gated slice-based imaging, normal patient breathing (no breath hold), and no contrast agents to enhance structures of interest. With this protocol the kidneys could experience significant displacement during the time interval between slice acquisition due to respiratory motion, resulting in jagged contours in sagittal and coronal planes. In addition, partial volume and motion artifacts combined to cause the poles to be poorly visualized or spuriously extended or foreshortened [28] (Figure 3).

The efficient object representation of m-reps offers the advantage that relatively small numbers of training images are required [20]. The number of training images for this study was estimated from pilot studies

to be 40-80 images; a total of 53 images were used to for the right kidney and 51 images were used for the left kidney. The training images were selected from a collection of 60 diagnostic CT images acquired using a liver imaging protocol that did not involve contrast material. Motion artifacts seen in the target images were minimal, resulting in a model that resisted deformations that would capture the motion artifacts seen in Figure 3.

Segmentation procedures

Two experienced humans (observers A and B in the Results section) defined the target kidneys slice-by-slice on the original image data using interactive region fill together with pixel-painting editing tools for fine sculpting [29]. This method was selected to force the users to make pixel-level decisions at every location on the boundary. The work was performed without time constraints over multiple sessions scheduled at the convenience of the participants. Although no formal statistical comparison was performed, anecdotally this procedure resulted in higher quality segmentations than contours generated under clinical conditions, which generally approximate the kidney boundary as contours composed of many straight-line segments much longer than the dimension of a pixel and thus do not fully capture pixel-scale boundary detail. For comparison with m-reps the set of 2D contours for each human segmentation were converted to a binary image and from that into a 3D tiled surface using marching cubes [30]. The small scale scalloping produced by pixel painting (Figures 4-5) were smoothed in the tiling process (Figure 6) and played little role in the final comparisons.

The target images were resampled using tri-linear interpolation to 0.2 cm x 0.2 cm x 0.2 cm for m-rep segmentation. The first step using m-reps is to determine a starting point for the m-rep model in the target image. In the future this initialization step will be automatic but in this study was performed by a graduate student who had no prior segmentation experience. This step involved interactively dragging and dropping the m-rep over the kidney to be segmented. A single soft-tissue intensity window was used for all target images. The segmented kidneys were produced in the form of 3D tiled surfaces that could be directly compared to the tiled surfaces computed from the hand-drawn 2D contours. Surface comparisons were performed using tools provided in VALMET [31].

4. Results

4.1. Example segmentations

Results of the best and worst segmentations, based on the metrics described above, are illustrated in Figures 4-6. Figure 4 shows good agreement in adjacent transverse slices of the kidney for the best case. Results near the midsection are shown because in the transverse plane disagreement tended to be more pronounced near the renal pelvis due to “structure noise” of tubular structures entering and exiting the renal pelvis. Figure 5 shows adjacent transverse slices through the midsection for the worst case. The region of disagreement in Figure 5 demonstrates a large change in shape from one slice to the next for the human observers. Such a large change would be resisted by the m-rep model used in this study, resulting in a smooth 3D surface through this region as seen in the left panel Figure 6.

4.2 Statistical analysis of distance separation between surfaces

Distance separation was examined by comparing segmented surfaces in pairs. The surfaces were designated as reference and trial, with each surface playing both roles. Histograms were built from measurements of the shortest distance between a point on the trial surface to the nearest point on the reference surface for 2562 points. This measurement suffers because it is not symmetric due to the lack of point correspondence between the two compared surfaces, a general problem that is not unique to this study. In particular for any point selected on a kidney surface produced by m-reps, the corresponding point is not uniquely defined on the surface of the same kidney produced by a human segmenter, and vice-versa. This lack of correspondence leads to asymmetry when measuring the distance between two surfaces. For example, the

distance from a point on the trial surface to the nearest point on the reference surface is not the same when measured in reverse (Figure 7). The approach chosen to deal with this problem was to measure distances between each pair of surfaces twice, with the role of reference and trial exchanged. The two resulting histograms were pooled by summing counts in individual distance bins. Two metrics derived from the distance histograms, Mean and Q4, were used to compare m-reps (denoted as segmenter "C") with human segmenters (referred to as "A" and "B", respectively). The mean is the average absolute distance over all test cases for a pair of segmenters, and Q4 is the fourth quartile of distances and is equivalent to the Hausdorff maximum separation distance. (Note: Quartile ratings give the surface separation associated with each quartile, e.g., a value of .18 cm for Q2 means that 50% of all points on the compared surfaces are separated by no more than 0.18 cm. In this study Q1-Q3 produced no discrimination between human-human and human-m-reps comparisons.)

Percent volume overlap can be defined several ways depending on the reference volume. In this study overlap was defined as the intersection of two segmentations divided by their union. Excluding the rare exception, which did not occur in this study, where one segmentation is contained entirely within the other, the union volume will be larger than either of the compared volumes. This results in smaller overlaps compared with using one, or the average, of the two segmentations as the reference (Table 1). For example in this study the reported (min, max) ranges for human-human and m-reps-human overlap were (92.6, 80.3) and (88.4, 76.8) respectively. These ranges increase to (96, 90) and (96, 84) when the average volume is used as the reference.

Table 1 displays the mean, Q4, and volume overlap with standard deviations for each segmenter pair over right and left kidneys grouped separately and together. Averaged over all kidneys the mean volume overlap for human segmentations was 88.8%, the mean surface separation was 0.12 cm, and the mean Hausdorff distance was 0.99 cm. The mean volume overlap between human and m-rep segmentations was 82-83%, the mean surface separation was 0.18-0.19 cm, and the mean Hausdorff distance was 1.14-1.25 cm. These results show that the two human observers compared slightly better with each other than with m-reps. As discussed in greater detail in Conclusions, these results are to be expected.

Repeated measures analysis of variance (REPM ANOVA) was performed to test each outcome (Mean, Q4, and Overlap). All tests were conducted at the same step-down level ($= 0.01$). Tests were performed for Side x Pair interaction, main effect of Pair, and main effect of Side. Table 2 reports p-values for these tests. Most tests were insignificant. The exceptions were Mean and Overlap, where the main effect of Pair was significant, with significant differences for AC/AB and BC/AB. Hence the distance between the two human segmentations was different from the distance of the m-reps segmentation to either human segmentation. Overall, mean±st.dev. distances (cm) were $\{0.12\pm 0.04, 0.18\pm 0.05, 0.19\pm 0.06\}$ for $\{AB, AC, BC\}$. Similarly, mean volume overlap (%) were $\{88.8\pm 0.82, 83.0\pm 1.46, 82.0\pm 1.45\}$.

5.0 Conclusions

Overall in this study the best m-rep kidney segmentations were at least as good as careful manual slice-by-slice segmentations, and the worst performance was probably no worse than humans under typical clinical conditions. Moreover m-rep performance was robust against the strong imaging artifacts present in the target images.

The mean surface separations between human and m-rep segmentations were slightly larger than for human-human segmentations but still in the sub-voxel range. Volume overlap and maximum surface separation also were slightly better for human-human comparisons. These results are not surprising since several factors in this study favored human-human comparison. The origins of disagreement can be grouped into four general classes, only one of which is related to the particular m-rep model used in this study. The areas of disagreement are 1) systematic differences between manual 2D and automatic 3D segmentation. Manual contouring produces a slab for each slice. As seen in Figure 8 slice-by-slice contouring created slabs that when joined together

resulted in 3D kidneys with stair-steps, while the m-rep model in this study produced smooth surfaces. The correspondence of the stair-steps in the segmentations of both humans and their total absence in the m-reps segmentations favored human-human comparison.; 2) imaging artifacts, e.g., motion due to breathing. Motion artifacts cause cross-sections of the same objects to be displaced in the transverse plane from slice to slice, generating more and wider stair-steps in the 3D surface created from stacked slabs. Objects also can be elongated and foreshortened. Slice-by-slice contouring tends to preserve imaging artifacts while m-rep segmentation has a smoothing effect.; 3) image voxel dimensions. In regions of high contrast in ideal images the inter-observer agreement for localizing an edge at the voxel level is limited primarily by the voxel dimensions. Poor contrast will degrade the level of agreement. Figure 6 illustrates that agreement can be quite good when voxel size is the main limiting factor.; and 4) the use of a single-figure m-rep that yielded a smooth surface across the renal pelvis. As seen in Figures 8 and 9, humans sometimes drew indentations at the renal pelvis. For those cases where both humans indented the absence of indentations in m-reps segmentations resulted in worse metrics for human-m-reps than for human-human comparisons.

Even though m-reps compared favorably with humans in this study a number of improvements and extensions are being investigated [21]. Improvements related to kidney models include developing intensity profiles that account for absolute intensity as well as relative shape in the image match term; developing a method for considering a mix of weighted intensity profiles instead of a single intensity profile at surface points during intensity training; developing a multi-figure model with an indentation to exclude structures in the renal pelvis from the segmented kidney; and developing a boundary level deformation stage that is robust against image disturbances.

Figures

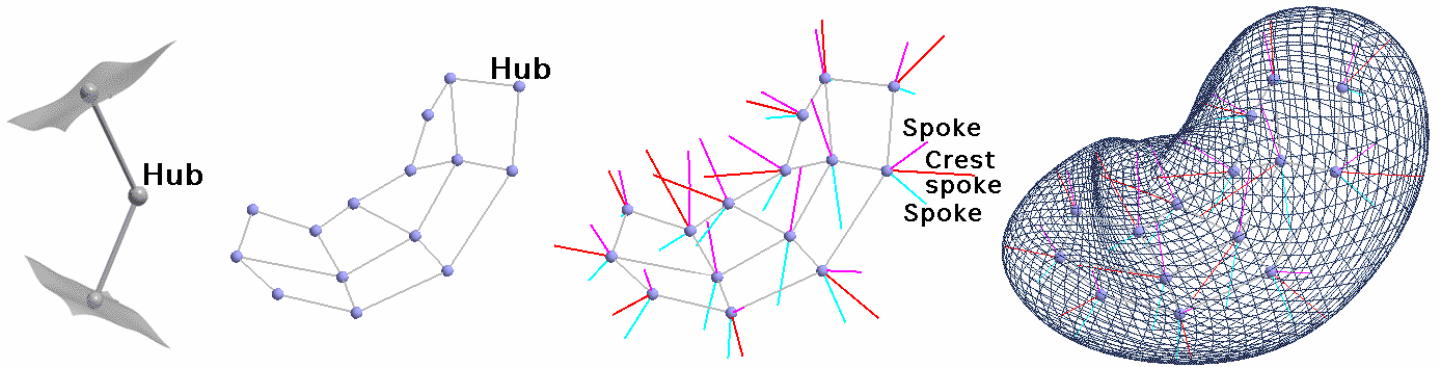


Figure 1. Frame 1: Medial atom with two equal-length spokes that touch points on surface patches on opposite sides of the object and thus define object width at the location of the atom. **Frame 2:** A medial sheet of a kidney as viewed from an oblique angle. The sheet is represented as a 5x3 grid of medial atoms with only the atom hubs displayed. **Frame 3:** Medial grid with spokes displayed. Internal atoms have two spokes (magenta and cyan) and atoms on the edge of the grid have a third spoke (red) that defines the radius of curvature of the crest of the object. **Frame 4:** Wire-frame rendering of the surface implied by the medial sheet.

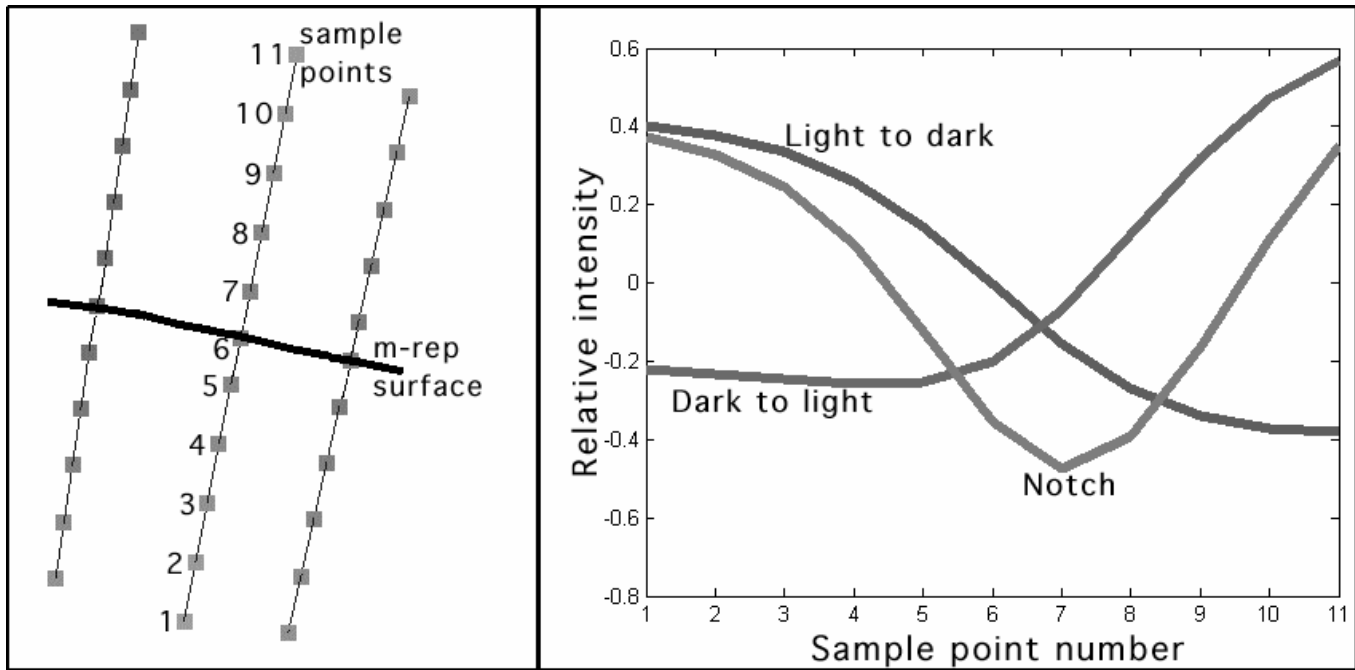


Figure 2. Left: Line segments for intensity training. The segments are perpendicular to the m-rep surface with the midpoint positioned on the surface. Intensity values are sampled at eleven evenly spaced points. **Right:** The three canonical forms for classifying intensity profiles.

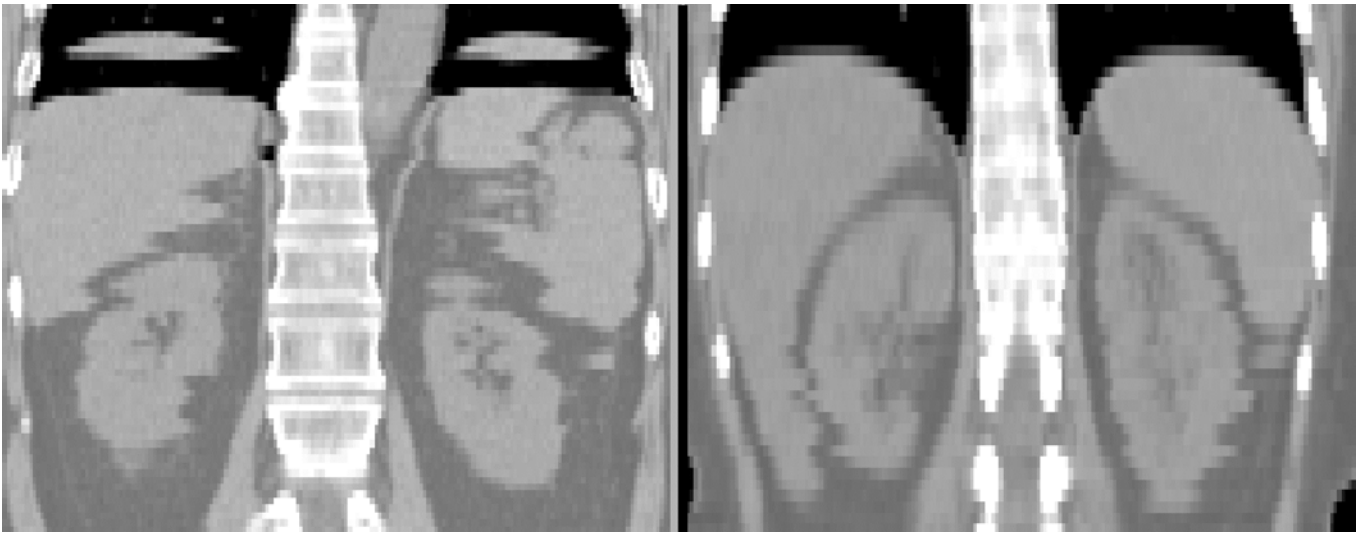


Figure 3. Coronal slices through two target images showing significant motion artifacts. In both images, adjacent slices of the kidneys are displaced in the transverse plane, and polar regions show signs of elongation and perhaps contraction. Slice-by-slice segmentation tends to preserve such artifacts.

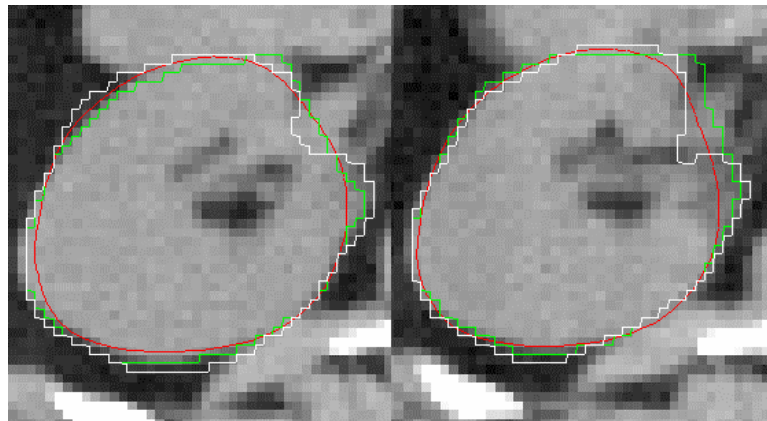


Figure 4. Adjacent transverse slices through the mid section of the kidney for the best case. The human segmentations are colored white and green, and m-reps is red.

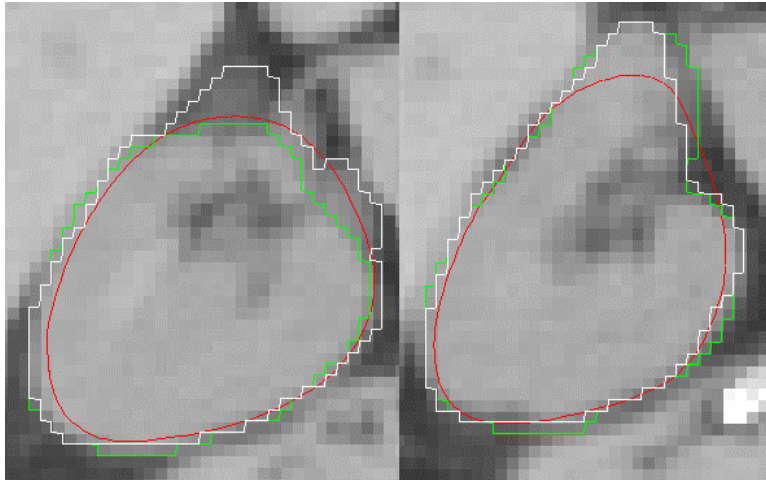


Figure 5. Adjacent transverse slices through the mid section of the kidney for the worst case. The human segmentations are colored white and green, and m-reps is red.

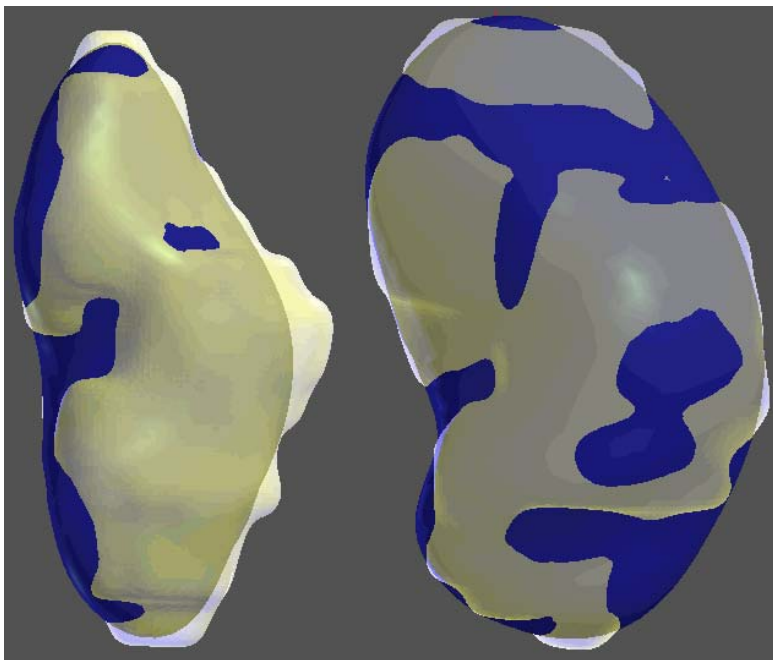


Figure 6. Left: Surface renderings for the worst case. The m-reps result is shown as a solid blue surface and the human segmentation is a white transparent surface. Notice the smooth m-rep surface near the region of disagreement at the midsection seen in Figure 5. Regions of disagreement appear to be associated primarily with the types of motion artifacts seen in Figure 2. **Right:** Surface renderings for the best case showing good agreement between human and m-rep segmentations, primarily because the image was relatively free of motion artifacts. Surface displacement is in the subvoxel range and thus related to image resolution.

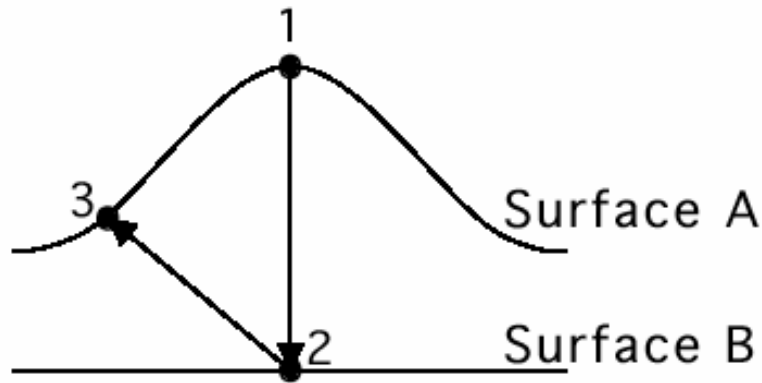


Figure 7. Illustration of the lack of symmetry when computing the minimum distance between two surfaces in this study. The minimum distance to surface B from Point 1 on surface A is defined by the line connecting points 1 and 2. However the minimum distance to surface A from Point 2 is defined by the line connecting points 2 and 3.

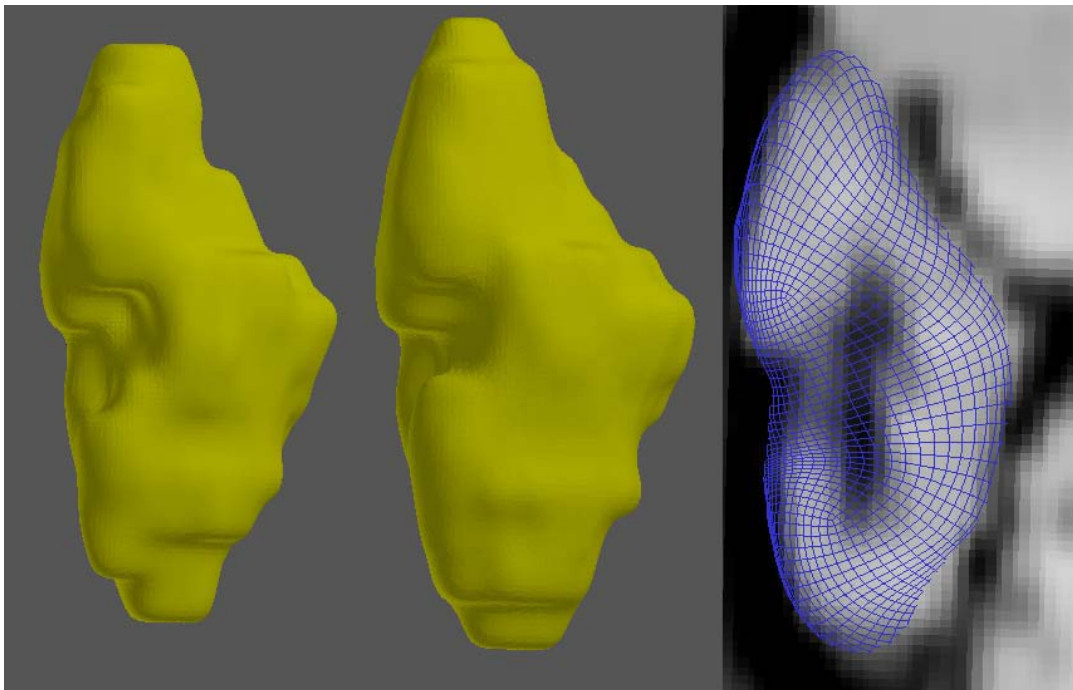


Figure 8. Left and Center: Surface renderings of the human segmentations for the worst case demonstrating inherent stair-steps that are exacerbated by motion artifacts. The center kidney demonstrates extra slabs at the top and bottom that also can result from motion artifacts. The segmentation on the left ignored the artifacts on these slices. The renal pelvis is indented for both segmentations. **Right:** Wire-frame rendering of the m-rep segmentation for the same case superimposed on the image data. The motion artifacts responsible for the stair-steps in the human segmentations are clearly visible in the image data. Note that m-reps resisted deformations that resulted in large changes from slice to slice.

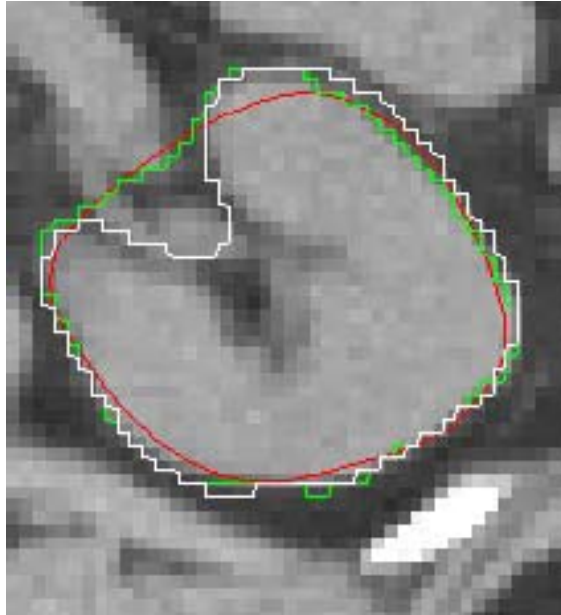


Figure 9. Slice through the renal pelvis showing how humans can differ in the way they deal with structures in this region. One human (green) contoured straight across the pelvis while the other (white) excluded some of the pelvic structures. Single-figure m-reps (red) produce a flat surface across the pelvis.

Tables

Table 1

Mean distance separation (Mean), Hausdorff or maximum separation distance (Q4), and volume overlap (Overlap) for human-human (AB), and human-m-reps (AC and BC) segmentations. Maximum and minimum values (Max, Min) are also given for each metric.

<i>Side</i>	<i>Pair</i>	<i>Mean (cm)</i>	<i>Q4 (cm)</i>	<i>Overlap(%)</i>
		<i>Max, Min(cm)</i>	<i>Max, Min(cm)</i>	<i>Max, Min(%)*</i>
Left	AB	0.11±0.03	1.03±0.35	88.8 ± 3.21
		0.19, 0.07	1.56, 0.57	92.5, 81.3
Left	AC	0.17±0.05	1.33±0.44	83.9 ± 5.41
		0.27, 0.10	2.19, 0.59	88.9, 78.7
Left	BC	0.18±0.07	1.13±0.48	83.1 ± 6.22
		0.33, 0.11	1.75, 0.49	87.8, 78.2
Right	AB	0.12±0.05	0.95±0.33	88.7 ± 4.17
		0.21, 0.07	1.64, 0.59	92.6, 80.3
Right	AC	0.19±0.06	1.18±0.34	82.0 ± 5.67
		0.30, 0.09	1.70, 0.68	87.8, 75.3
Right	BC	0.20±0.05	1.16±0.29	80.9 ± 5.01
		0.27, 0.09	1.67, 0.78	88.4, 76.8
Both	AB	0.12±0.04	0.99±0.34	88.8±0.82
		0.21,0.07	1.64, 0.57	92.6, 80.3
Both	AC	0.18±0.05	1.25±0.39	83.0±1.46
		0.30, 0.09	2.19, 0.59	88.9, 75.3
Both	BC	0.19±0.06	1.14±0.39	82.0±1.45
		0.33, 0.09	1.75, 0.49	88.4, 76.8

*Choosing the union results in smaller overlaps compared to choosing one, or the average, of the compared volumes as the reference. In this study the (max, min) ranges for overlap are (96, 90) and (96, 84) for human-human and human-m-reps respectively when the average volume is used as the reference.

Table 2

P-values for statistical tests for interactions.

	Mean	Q4	Overlap
Side x Pair interaction	0.4415	0.2423	0.4249
Pair main effect	0.0052	0.1063	0.0100
AC / AB	0.0074	0.0336	0.0100
BC / AB	0.0010	0.2064	0.0022
BC / AC	0.1869	0.2107	0.2899
Side main effect	0.2746	0.5924	0.2424

References

1. Dowsett RJ, Galvin JM, Cheng E, et al. Contouring structures for 3-dimensional treatment planning. *Int J Radiat Oncol Biol Phys* 1992; 22:1083-1088.
2. Leunens G, Menten J, Weltens C, et al. Quality Assessment of Medical Decision Making in Radiation Oncology: Variability in Target Volume Delineation for Brain Tumors. *Radiother and Oncol* 1993; 29:169-175.
3. Valley J-F and Mirimanoff R-O. Comparison of treatment techniques for lung cancer. *Radiother and Oncol* 1993; 28:168-173.
4. Kagawa K, Lee WR, Schultheiss TE, et al. Initial clinical assessment of CT-MRI image fusion software in localization of the prostate for 3D conformal radiation therapy, *Int J Radiat Oncol Biol Phys* 1997; 38: 319-325.
5. Roach M, Akazawa PF, Malfatti C, et al: Prostate volumes defined by magnetic resonance imaging and computerized tomographic scans for three-dimensional conformal radiotherapy. *Int J Radiat Oncol Biol Phys* 1996; 35: 1011-1018.
6. Algan O, Hanks GE, and Shaer AH. Localization of the prostatic apex for radiation treatment planning, *Int J Radiat Oncol Biol Phys* 1995; 33: 925-930.
7. Ketting C, Austin-Seymour M, Kalet I, et al. Consistency of Three-Dimensional Planning Target Volumes Across Physicians and Institutions. *Int J Radiat Oncol Biol Phys* 1997a; 37: 445-453.
8. Ketting C, Austin-Seymour M, Kalet I, et al. Automated planning target volume generation: An evaluation pitting a computer-based tool against human experts. *Int J Radiat Oncol Biol Phys* 1997b; 37: 697-704.
9. Rasch C, Barillot I, Remeijer P, et al. Definition of the prostate in CT and MRI: A multi-observer study. *Int J Radiat Oncol Biol Phys* 1999; 43: 57-66.
10. Kass M, Witkin A, Terzopoulos D. Snakes: Active Contour Models. *Int J Comp Vision* 1987; 1:321-331.
11. ter Haar Romeny BM (Ed): *Geometry-Driven Diffusion in Computer Vision*, 1994, Kluwer Academic Press.
12. CVRMed '95: Proc First Int Conf CVRMed, *Lecture Notes in Computer Science* 1995; 905, N Ayache (Ed), Springer.
13. CVRMed-MRCAS '97: Proc 1st Joint Conference Computer Vision, Virtual Reality and Robotics in Medicine and Medical Robotics and Computer-Assisted Surgery, *Lecture Notes in Computer Sciences* 1997;. 1205, Troccaz, J, Grimson E, and Moesges R (Eds), Springer.
14. Montagnat J and Delingette H. Volumetric medical images segmentation using shape constrained deformable models. In: *Lecture Notes in Computer Sciences* 1997; 1205:13-22, Troccaz, J., Grimson, E., and Moesges, R. (eds), Springer.
15. McInerney T and Terzopoulos D. Deformable Models in Medical Image Analysis. *Proc Workshop on Math*

Methods in Biomed Imag Anal 1996a; IEEE Cat. # 96TB100056: 171-180.

16. McNerny T, Terzopoulos D. Deformable models in medical image analysis: a survey. *Medical Image Analysis* 1996b; 1(2): 91-108.

17. Jones TN and Metaxas DN: Segmentation using models with affinity-based localization. In: *Lecture Notes in Computer Sciences* 1997; 1205: 53-62, Troccaz, J., Grimson, E., and Moesges, R. (eds), Springer.

18. Vehkomäki T, Gerig G, and Székely GA. User-guided tool for efficient segmentation of medical image data. In: *Lecture Notes in Computer Sciences* 1997; 1205: 685-694, Troccaz, J., Grimson, E., and Moesges, R. (eds), Springer.

19. *IEEE TMI Special Issue on Model-Based Analysis of Medical Images*, 18(10) (1999).

20. Pizer SM, Fletcher PT, Joshi S, et al. Deformable m-reps for 3D medical image segmentation. *Int J Comp Vision* 2003; 55:85-106.

21. Pizer SM, Fletcher PT, Joshi S, et al. A method & software for segmentation of anatomic object ensembles by deformable m-reps. *Med Phys* 2004. (Submitted)

22. Fletcher PT, Joshi S, Lu C et al. Principal Geodesic Analysis for the Study of Nonlinear Statistics of Shape. *IEEE Trans Med Imag* 2004 (in press, available at <http://midag.cs.unc.edu/pubs/papers/>).

23. Lu C, Pizer S, Joshi S. A Markov Random Field Approach to Multi-scale Shape Analysis. Scale Space Methods in Computer Vision, LD Griffin and M Lillholm, eds. *Lecture Notes in Computer Science* 2003; 2695: 416-431.

24. Pizer SM, Medial & Medical: A Good Match for Image Analysis. *Int J Comp Vis* 2003; 55: 79-84.

25. Styner M, Gerig G, Pizer S, et al. Automatic and Robust Computation of 3D Medial Models Incorporating Object Variability. *Int J Comp Vis* 2003; 55:107-122.

26. Stough J, Pizer SM, Chaney EL et al. Clustering on Image Boundary Regions for Deformable Model Segmentation. *Proc Int Symp Biomed Imag* 2004; IEEE Cat No 04EX821C: 436-439. Available at <http://midag.cs.unc.edu/pubs/papers/>.

27. Thall A. Fast C^2 interpolating subdivision surfaces using iterative inversion of stationary subdivision rules. UNC Dept Comp Sci Tech Report TR02-001, 2003. Available at <http://midag.cs.unc.edu/pubs/papers/>.

28. Pan T, Lee TY, Rietzel E, Chen GT. 4D-CT imaging of a volume influenced by respiratory motion on multi-slice CT. *Med Phys* 2004; 31(2): 333-340.

29. Tracton G, Chaney EL, Rosenman JG, et al. MASK: Combining 2D and 3D Segmentation Methods to Enhance Functionality. *Math Methods Med Imaging III*, SPIE 1994; 2299: 98-109.

30. Lorensen WE and Cline HE. Marching Cubes: A High Resolution 3D Surface Construction Algorithm, *Computer Graphics* 1987; 21(4):163-169.
31. Gerig G, Jomier M, Chakos M. Valmet: A new validation tool for assessing and improving 3D object segmentation. *Proc MICCAI 2001*; Springer LNCS 2208: 516-523.