Int. J. Radiation Oncology Biol. Phys., Vol. 61, No. 3, pp. 954–960, 2005

0360-3016/05/$–see front matter

doi:10.1016/j.ijrobp.2004.11.014

PHYSICS CONTRIBUTION

COMPARISON OF HUMAN AND AUTOMATIC SEGMENTATIONS OF KIDNEYS FROM CT IMAGES

MANJORI RAO, M.S.,* JOSHUA STOUGH, B.S.,† YUEH-YUN CHI, M.S.,‡ KEITH MULLER, PH.D.,‡ GREGG TRACTON, B.S.,* STEPHEN M. PIZER, PH.D.,*† AND EDWARD L. CHANEY, PH.D.*

Departments of *Radiation Oncology, †Computer Science, and ‡Biostatistics, University of North Carolina, Chapel Hill, NC

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 computerized tomographic (CT) images.

Methods and Materials: Deformable shape models called m-reps were used to automatically segment right and left kidneys from 12 target CT images, and the results were compared with 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, the medial atoms comprising the m-rep model were deformed one by one. This procedure was iterated until no changes were observed. The transformations and deformations at both stages were driven by optimizing an objective function with two terms. One term penalized the currently deformed m-rep by an amount proportional to its deviation from the mean m-rep 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 article 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 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%.

Conclusions: Overall in this study, the best m-rep kidney segmentations were at least as good as careful manual slice-by-slice segmentations performed by two experienced humans, and the worst performance was no worse than typical segmentations from our clinical setting. The mean surface separations for human–m-rep segmentations were slightly larger than for human–human segmentations but still in the subvoxel 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. © 2005 Elsevier Inc.

Image segmentation, Kidney, Treatment planning.

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 isocenters 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 anatomic structures and tumor-related objects comprising the patient

Reprint requests to: Edward L. Chaney, Ph.D., Department of Radiation Oncology, University of North Carolina, Campus Box 7512, 101 Manning Dr, Chapel Hill, NC 27599-7512. Tel: (919) 966-0300; Fax: (919) 966-7681; E-mail: chaney@med.unc.edu

This research was supported by NCI P01 EB002779.

Received Jun 25, 2004, and in revised form Oct 25, 2004. Accepted for publication Nov 1, 2004.

model are defined by segmenting one or more volume images, usually computerized tomographic (CT) and magnetic resonance 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 suboptimal treatment planning include intra- and interuser variabilities (1–9), 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 three-dimensional (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 to steer user-directed planning decisions and inverse treatment planning algorithms consistently in the right direction. Deformable shape models are a general class that is showing great promise for automatic segmentation of normal anatomic structures. Kass *et al.* (10) first described a straightforward method based on deformable two-dimensional contours popularly known as snakes. A useful survey of snakes is found in the study by McInerney and Terzopoulos (11). Collections of articles on early deformable models can be found in the book by ter Haar Romeny (12) and in proceedings of conferences such as CVRMed '95 (13) and CVRMed-MRCAS '97 (14); the topic is also investigated in studies by Montagnat and Delingette (15), McInerney and Terzopoulos (11, 16), Jones and Metaxas (17), and Vehkomäki *et al.* (18). 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. A special issue of the Institute of Electrical and Electronics Engineers' (IEEE) journal *Transactions on Medical Imaging* (19) on model-based analysis of medical images has a collection of articles 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 anatomic 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 *et al.* (20, 21), Fletcher *et al.* (22), and Lu *et al.* (23).

In this article, we discuss the results of an observer study comparing automatic and human segmentations of left and right kidneys from planning CT images. The objective was to compare m-reps against experienced humans to judge whether m-reps produce reasonable segmentations. To accomplish this, we conducted a biostatistically rigorous comparison of m-reps against two exemplars from the population of experienced humans. 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).

METHODS AND MATERIALS

M-reps

Detailed discussions of the structure, building, training, and deformation of m-reps can be found in articles by Pizer *et al.* (20, 21). For completeness and continuity, brief discussions relevant to the kidney m-reps used in this study are presented below.

The simplest 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. Such an object is described using an m-rep model comprising a grid of atoms that implies a 3D surface, as shown in Fig. 1. The centers of

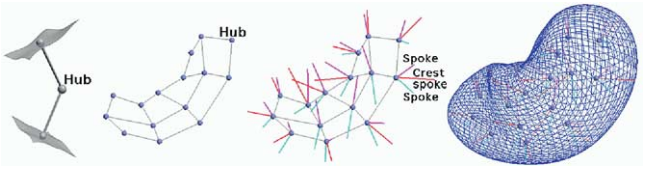

Fig. 1. Frame 1: Medial atom with two equal-length spokes that touch points on surface patches on opposite sides of the object and thus define object width at the location of the atom. Frame 2: A medial sheet of a kidney as viewed from an oblique angle. The sheet is represented as a 5 × 3 grid of medial 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.

the atoms, called hubs, lie on the medial sheet. Interior atoms have two spokes of equal length that extend to patches on opposite sides of the implied surface. Edge atoms have a third spoke defining the radius of curvature for the crest section of the implied surface. The number of atoms in an m-rep can be selected to be the fewest needed to capture the full range of shape variability over the target population (25). A 5 × 3 grid was used in this study (Fig. 1).

M-reps are trained using a set of images representing the population of interest. Truth is defined in the training images by experienced humans who segment the kidneys using slice-by-slice contouring. 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). 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 2,562 points over the m-rep surface (26–28). The relative intensity variation at each location, called an intensity profile, was measured along line segments that passed through the points (Fig. 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 point was measured across all training images, and the resulting collection was compared with 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 and the like; (*2*) dark to light, capturing kidney boundary locations abutting lighter liver, bone, and other structures; and (*3*) a notch, capturing kidney boundary locations with a small amount of darker fat between the kidney and another section of organ tissue or bone. For this study, the characteristic profile identified with a particular point 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. The segmentation algorithm runs on a modern personal computer under the Windows operating system. The segmentation time per kidney is on the order of 1–3 min using the current version on our research software, which has not been optimized for speed. 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 are linear combinations of the principal modes of variation determined by PGA. Atom-by-atom deformations define the next scale. 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.

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 determined by PGA. 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.

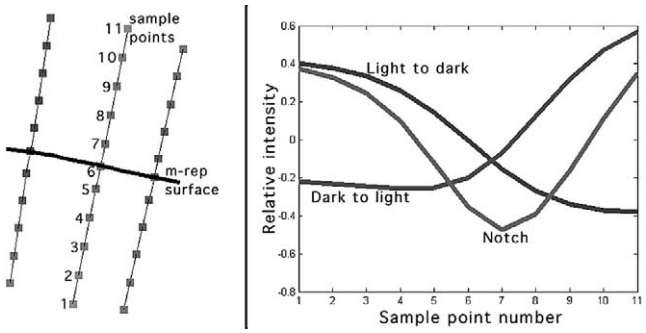

Fig. 2. Left: Line segments for intensity training. The segments are perpendicular to the m-rep surface with the midpoint positioned on the surface. Intensity values are sampled at 11 evenly spaced points. Right: The three canonical forms for classifying intensity profiles.

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 × 0.098 mm to 0.156 × 0.156 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 nongated 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) (Fig. 3). Image artifacts can obscure "truth," and thus comparison with humans in localized regions affected by artifacts is ambiguous. Because the intent was to evaluate m-reps on actual planning images, however, the challenge posed by the motion artifacts had to be accepted.

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 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

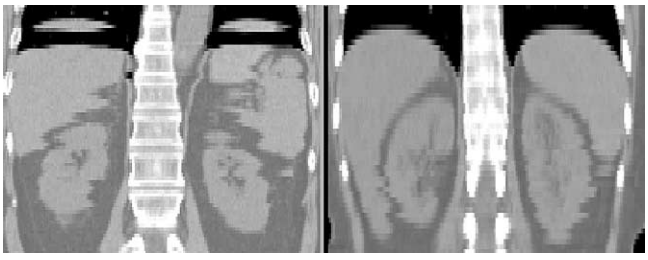

Fig. 3. Coronal slices through two target images showing significant motion artifacts. In both images, adjacent slices of the kidneys are displaced in the transverse plane, and polar regions show signs of elongation and perhaps contraction. Slice-by-slice segmentation tends to preserve such artifacts.

material. Motion artifacts in the training images were minimal, resulting in a model that resisted deformations that would capture the motion artifacts seen in Fig. 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 two-dimensional contours for each human segmentation was 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 (Figs. 4–5) was smoothed in the tiling process (Fig. 6) and played little role in the final comparisons.

The target images were resampled using tri-linear interpolation to 0.2 cm × 0.2 cm × 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 it 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 with the tiled surfaces computed from the hand-drawn two-dimensional contours. Surface comparisons were performed using tools provided in Valmet (31).

RESULTS

Example segmentations

Results of the best and worst segmentations, based on the metrics described earlier, are illustrated in Figs. 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 owing 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 Fig. 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 of Fig. 6.

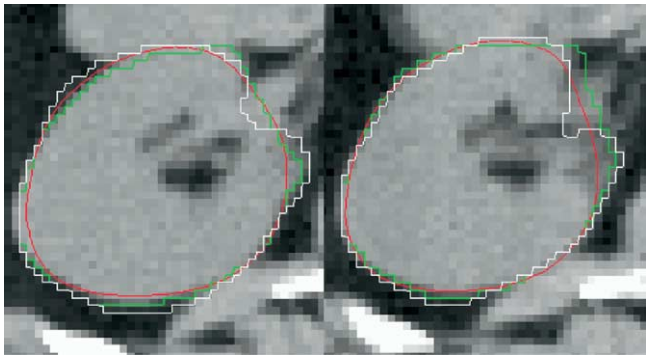

Fig. 4. Adjacent transverse slices through the midsection of the kidney for the best case. The human segmentations are colored white and green, and m-reps are red.

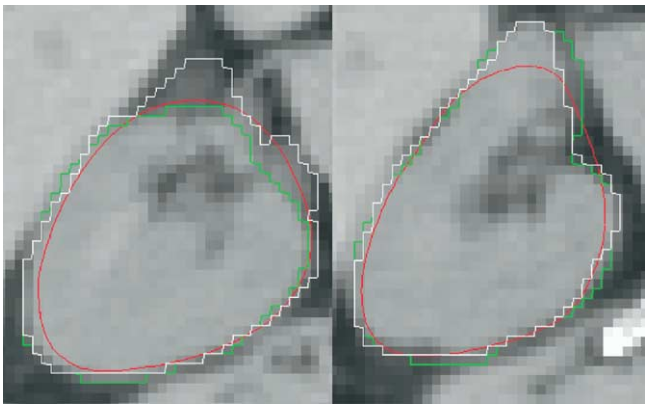

Fig. 5. Adjacent transverse slices through the midsection of the kidney for the worst case. The human segmentations are colored white and green, and m-reps are red.

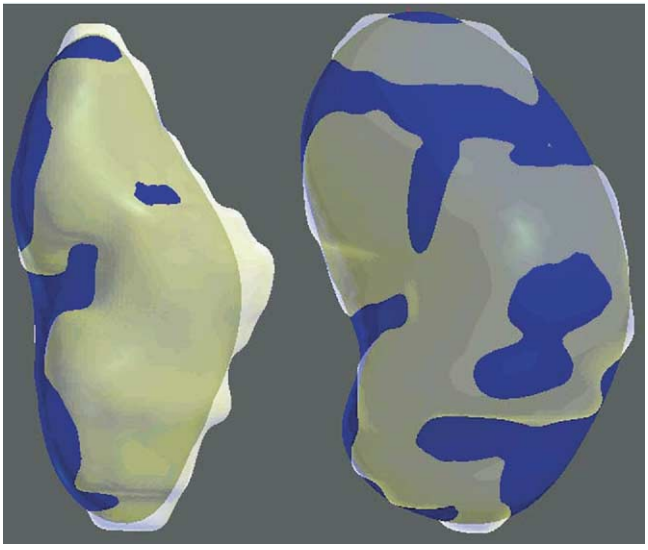

Fig. 6. Left: Surface renderings for the worst case. The m-reps result is shown as a solid blue surface, and the human segmentation is a white transparent surface. Notice the smooth m-rep surface near the region of disagreement at the midsection seen in Fig. 5. Regions of disagreement appear to be associated primarily with the types of motion artifacts seen in Fig. 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.

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 2,562 points. This measurement suffers because it is not symmetric, as a result of 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 (Fig. 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 0.18 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 in 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 1, 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.

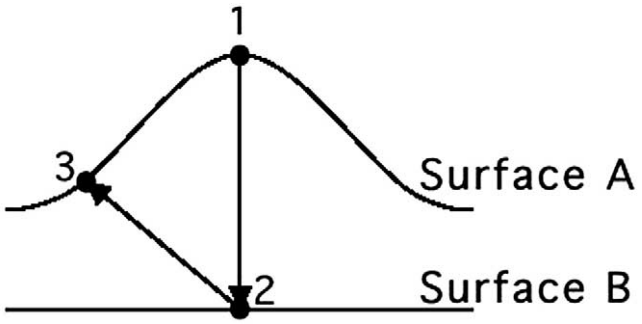

Fig. 7. Illustration of the lack of symmetry when computing the minimum distance between two surfaces in this study. The minimum distance to surface B from point 1 on surface A is defined by the line connecting points 1 and 2. However, the minimum distance to surface A from point 2 is defined by the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

Repeated measures analysis of variance 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 × 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

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

Side	Pair	Mean (cm) Max, Min (cm)	Q4 (cm) Max, Min (cm)	Overlap (%) Max, Min (%)*
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 with 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.

from the distance of the m-reps segmentation to either human segmentation. Overall, mean ± SD distances (cm) were {0.12 ± 0.04, 0.18 ± 0.05, 0.19 ± 0.06} for {AB, AC, BC}. Similarly, mean volume overlap (%) was {88.8 ± 0.82, 83.0 ± 1.46, 82.0 ± 1.45}.

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 in our clinical setting. 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 subvoxel range. Volume overlap and maximum surface separation also were slightly better for human–human comparisons. These results are not surprising, because 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 as follows: (*1*) Systematic differences between manual two-dimensional and automatic 3D segmentation. Manual contouring produces a slab for each slice. As seen in Fig. 8 slice-by-slice contouring created slabs that, when joined together, resulted in 3D kidneys with stair-steps, whereas 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, whereas m-rep segmentation has a smoothing effect. (*3*) Image voxel dimensions. In regions of high contrast in ideal images the interobserver 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. (*4*) The use of a single-figure m-rep that was trained to produce a smooth surface across the renal pelvis. As seen in Figs. 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.

Table 2. *P* values for statistical tests for interactions

	Mean	Q4	Overlap
Side × 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

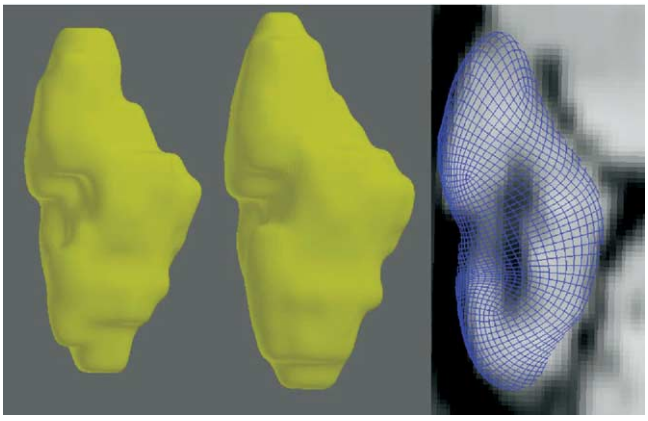

Fig. 8. Left and center: Surface renderings of the human segmentations for the worst case demonstrating inherent stair-steps that are exacerbated by motion artifacts. The center kidney demonstrates extra slabs at the top and bottom that also can result from motion artifacts. The segmentation on the left 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.

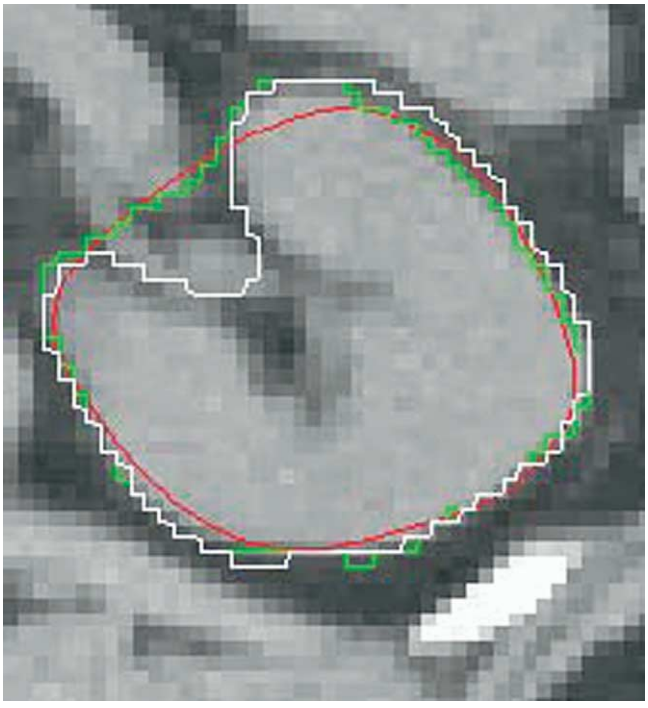

Fig. 9. Slice through the renal pelvis showing how humans can differ in the way they deal with structures in this region. One human (green) contoured straight across the pelvis, whereas the other (white) excluded some of the pelvic structures. Single-figure m-reps (red) produce a flat surface across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

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 multifigure 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.

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 Oncol* 1993;29:169–175.
3. Valley J-F, Mirimanoff R-O. Comparison of treatment techniques for lung cancer. *Radiother 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, 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* 1997;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* 1997;37:697–704.
9. Rasch C, Barillot I, Remeijer P, *et al.* Definition of the prostate in CT and MRI: A multiobserver 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. McInerney T, Terzopoulos D. Deformable models in medical image analysis. Proceedings of a Workshop on Mathematical Methods in Biomedical Image Analysis, CAT 96TB100056, 1996: Los Alamitos, CA: Institute of Electrical and Electronics Engineers (IEEE); 1996; p. 171–180.
12. ter Haar Romeny BM, ed. Geometry-driven diffusion in computer vision. Dordrecht, The Netherlands: Kluwer Academic Press; 1994.
13. CVRMed '95: Proceedings of the First International Conference on Computer Vision, Visual Reality, and Robotics in Medicine. In: Ayache N, editor. *Lecture Notes in Computer Science*. New York: Springer; 1995:905.
14. CVRMed-MRCAS '97: Proceedings of the 1st Joint Conference on Computer Vision, Virtual Reality and Robotics in Medicine and Medical Robotics and Computer-Assisted Surgery. In: Troccaz J, Grimson E, Moesges R, editors. *Lecture Notes in Computer Science*. New York: Springer; 1997;1205.
15. Montagnat J, Delingette H. Volumetric medical images segmentation using shape constrained deformable models. In: Troccaz J, Grimson E, Moesges R, editors. *Lecture Notes in Computer Science*. New York: Springer; 1997;1205:13–22.
16. McInerney T, Terzopoulos D. Deformable models in medical image analysis: a survey. *Med Image Analysis* 1996;1:91–108.
17. Jones TN, Metaxas DN. Segmentation using models with affinity-based localization. In: Troccaz J, Grimson E, Moesges R, editors. *Lecture Notes in Computer Science*. New York: Springer; 1997;1205:53–62.
18. Vehkomäki T, Gerig G, Székely GA. User-guided tool for efficient segmentation of medical image data. In: Troccaz J, Grimson E, Moesges R, editors. *Lecture Notes in Computer Science*. New York: Springer; 1997;1205:685–694.
19. Institute of Electrical and Electronics Engineers (IEEE). *Transactions on Medical Imaging*. Special issue on model-based analysis of medical images. 1999;18.
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 and software for segmentation of anatomic object ensembles by deformable m-reps. *Med Phys*. Accepted. Available at http://midag.cs.unc.edu/pubs/papers/.
22. Fletcher PT, Joshi S, Lu C, *et al.* Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans Med Imag*. Accepted. 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. In: Griffin LD, Lillholm M, editors. *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. University of North Carolina Department of Computer Science Technical Report TR02–001; 2003. Available at http://midag.cs.unc.edu/pubs/papers/.
28. Chen GTY, Kung JH, Beaudette KP. Artifacts in computed tomography scanning of moving objects. *Sem Rad Oncol* 2004;14:19–26.
29. Tracton G, Chaney EL, Rosenman JG, *et al.* MASK: Combining 2D and 3D segmentation methods to enhance functionality. Mathematical Methods in Medical Imaging III. Bellingham, WA: The International Society for Optical Engineering. 1994;2299:98–109.
30. Lorensen WE, Cline HE. Marching cubes: A high resolution 3D surface construction algorithm. *Computer Graphics* 1987; 21:163–169.
31. Gerig G, Jomier M, Chakos M. Valmet: A new validation tool for assessing and improving 3D object segmentation. Proceedings of the Fourth International Conference on Medical Image Computing and Computer-Assisted Intervention. *Lecture Notes in Computer Science*: New York: Springer. 2001;2208: 516–523.