PAPER

Construction of Abdominal Probabilistic Atlases and Their Value in Segmentation of Normal Organs in Abdominal CT Scans*

Hyunjin PARK[†], **Alfred HERO**[††], **Peyton BLAND**[†††], **Marc KESSLER**[††††], ***Nonmembers***,
Jongbum SEO[†††††], ***Member***, ***and*** **Charles MEYER**[†††], ***Nonmember***

SUMMARY A good abdominal probabilistic atlas can provide important information to guide segmentation and registration applications in the abdomen. Here we build and test probabilistic atlases using 24 abdominal CT scans with available expert manual segmentations. Atlases are built by picking a target and mapping other training scans onto that target and then summing the results into one probabilistic atlas. We improve our previous abdominal atlas by 1) choosing a least biased target as determined by a statistical tool, i.e. multidimensional scaling operating on bending energy, 2) using a better set of control points to model the deformation, and 3) using higher information content CT scans with visible internal liver structures. One atlas is built in the least biased target space and two atlases are built in other target spaces for performance comparisons. The value of an atlas is assessed based on the resulting segmentations; whichever atlas yields the best segmentation performance is considered the better atlas. We consider two segmentation methods of abdominal volumes after registration with the probabilistic atlas: 1) simple segmentation by atlas thresholding and 2) application of a Bayesian maximum a posteriori method. Using jackknifing we measure the atlas-augmented segmentation performance with respect to manual expert segmentation and show that the atlas built in the least biased target space yields better segmentation performance than atlases built in other target spaces.

key words: *atlas construction, segmentation, target selection, multidimensional scaling*

1. Introduction

The study of scans of a population leads to statistics of the population, which can be represented in a probabilistic atlas. The atlas typically contains information regarding shape and/or grayscale value variability of the population [1]. Probabilistic atlases have applications in segmentation [2]–[4] and registration [5], [6]. Atlases of the brain have been the most sought [6]. While the methodology to build a brain atlas can be applied to abdominal organs, few have actually built an abdominal atlas [7]–[9]. Previously we successfully built and published an abdominal probabilistic atlas from CT scans [4] consisting of liver, spinal cord, left kidney, and right kidney. Here we improve on the previously built abdominal atlas of CT scans by 1) choosing a least biased target scan among training scans to build an atlas, 2) choosing a better set of control points for the registration process, and 3) using CT scans having higher information content. We then apply segmentation algorithms to assess the quality of the atlas built. The resulting accuracy of the segmentation algorithm is used to assess the quality of the atlas. Previously our methodology to choose the target was tested on 2D simulated MRI [10]. Here we tested the target selection methodology based on the segmentation performance of real 3D CT scans.

Traditionally, researchers build their atlas by picking a target scan and mapping other training scans onto the target. Statistical processing can be performed on the same spatial frame after all scans are mapped onto the target. Statistical processing can be as simple as a grayscale average or some measure of probability at every voxel location. Methods for registration in terms of degrees of freedom (DOF) and geometric interpolant have to be the same for all registration tasks to ensure consistent construction and use of the atlas. Unfortunately, the resulting atlas is inherently biased by the selection of the chosen target scan because the atlas information is computed on the target's spatial frame. One way the bias towards a specific target may be reduced is by repeating the whole process of mapping other scans onto the target with the target replaced with an average scan from the previous registrations until the average scan converges [11]. Guimond et al. have shown the convergence rate of such iterative an approach [12].

Some researchers construct the atlas by registering all training scans at the same time [13]–[17]. In this approach, there is very little bias since the target space is very close to the mean geometry at the expense of increased computation complexity. Joshi et al. proposed a target free atlas construction method, but it has constraints on the geometric deformation it can handle [11]. Marsland et al. proposed to construct an atlas on a target scan that is close to the mean geometry of the training scans [18]. Our method of target selection in this paper shares a similar approach. We assume that an atlas is built by choosing a target scan and mapping other training scans onto the target. We choose the target scan which is the closest (i.e., least biased) to the mean geometry of the population using a well known statistical tool,

Manuscript received December 8, 2009.
Manuscript revised March 29, 2010.

[†]The author is with the Dept of Biomedical Engineering, Gachon Univ. of Medicine and Science, Incheon, Korea.
[††]The author is with the Dept of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA.
[†††]The authors are with the Dept of Radiology, University of Michigan, Ann Arbor, MI 48109, USA.
[††††]The author is with the Dept of Radiation Oncology, University of Michigan, Ann Arbor, MI 48109, USA.
[†††††]The author is with the Dept of Biomedical Engineering, Yonsei University, Wonju, Korea.

*This project was supported by the USPHS, National Institute of Health, NCI under grant 1P01CA87634 and Ministry of Knowledge and Economy, Korea.
DOI: 10.1587/transinf.E93.D.2291

multidimensional scaling (MDS).

Segmentation assigns labels to grayscale values or a contour that separates distinct regions. Outcome of a segmentation algorithm is a partitioned scan described by a small number of labels. Here we have a 5-label model where a CT scan gets segmented into liver, spinal cord, left kidney, right kidney, or "none of the above". Segmentation in itself is a vast field. We suggest the following review articles for a good overview of segmentation [19]. Here, we are particularly interested in segmentation algorithms using an atlas as side information. Atlas-based segmentation algorithms use atlas information as prior probabilities in a Bayesian framework or as a starting guess [20]–[23]. In this manner atlas information can guide segmentation algorithms where there is little discriminating grayscale information available. Since our focus in this paper is the value of atlas information for segmentation, we test our atlas construction method via two segmentation algorithms that use the atlas information. First, we consider a simple segmentation by registration approach; we register a test scan onto the atlas space and then threshold the probabilistic atlas to generate the segmentation of the test scan. Second, we consider a maximum a posteriori (MAP) segmentation where atlas information enters the formulation as a prior probability, which we implemented in our last paper [4].

We start with CT oncology scans of the abdomen whose expert manual segmentations by oncology therapy planners are available. We choose a target scan and then register other training scans onto that target. Once all scans are registered onto the target space, we apply the same mappings to the manual segmentations of the CT scans so that all manual segmentations are mapped onto the same target space. The organ probabilities in the atlas are computed from the summation of these mapped manual segmentations. For every voxel, we compute the probability of a voxel belonging to a certain organ. Depending on the chosen target, there may be many atlases constructed from the same set of training scans. One atlas is computed using the target space determined to be the best by MDS. Other atlases are computed using other target spaces. Then the value of an atlas is assessed based on the results from the atlas-based segmentation methods; whichever atlas yields the best segmentation performance is considered the best atlas.

The remainder of this paper is organized in the following manner. The first section covers the methods of atlas construction. The second section covers the methods of atlas-based segmentation. Finally, we conclude with the Discussion and Summary. The primary contributions of this paper are 1) to build a probabilistic atlas of abdominal organs using a least biased target and 2) to show the relative value of differently constructed atlases in simple atlas-based segmentation algorithms.

2. Methods; Atlas Construction

2.1 Registration Framework

Atlas construction involves many tasks of mapping one scan onto another scan. This task of mapping is called registration. Registration has been well reviewed in Hill's paper [24]. Basically two main components need to be addressed for any registration algorithm

- the similarity measure which measures degree of alignment, and
- the geometric interpolant which defines the geometric transform between two scans.

We choose mutual information (MI) as the similarity measure and thin-plate splines (TPS) as the geometric interpolant [25], [26]. Computing the MI involves calculating probability density functions of grayscale value distributions. A simple histogram with fixed bin width is used to calculate the probability density function. The process of registration can be formulated as maximizing the chosen similarity measure (i.e., MI) under a hypothetical geometric transform,

$$\hat{T} = \arg\max_{T \in F} MI(A(\bullet), B(T(\bullet))$$
$$\hat{T}; \text{estimate of the transform}$$
$$F; \text{family of feasible transforms} \quad (1)$$

A simplex optimizer is used to maximize the cost function [27]. One can choose other combinations of similarity measure (e.g., Normalized MI) and geometric transform (e.g., B-spline) [28].

The degrees of freedom (DOF) of TPS are determined by the number and locations of control points. Complex geometric transforms are modeled by many control points. TPS based registration requires the user to provide an initial guess of the transform (i.e., approximately specifying locations of control points in both scans). Initializing a TPS for high DOF is cumbersome. Thus, we employ a standard multi-level approach, where DOF is increased gradually. First, we start the registration process with 4 control points, which defines the affine transform, and then increase DOF (i.e., number of control points). Higher DOF registration is always initialized with the result from the previous lower DOF registration. As a result the registration process is automatic after the initial placement of 4 control points.

2.2 Control Points

With TPS the effect of control points is primarily, though not strictly, local. For example, control points in liver region primarily affect geometric transform in the liver region. This is particularly true for high DOF TPS transforms [4]. If one is interested in strictly local properties for the geometric transform, B-splines is a good choice. We employ a total of 43 control points where the liver has 24 control points, both

kidneys have 6 control points each, and the spinal cord has 7 control points. Within each organ, control points are approximately uniformly distributed in space similar to Fig. 1 of [4]. The liver is the largest organ among four organs modeled, which primarily drives the registration process. Hence it is important to allocate most control points to get the liver registration correct. We justify the use of 24 liver control points by the following. First, take two scans with available manual liver segmentations and mask them so that scans include liver and only its immediate vicinity. Perform registration of two scans with respect to varying number of control points of 4, 6, 12, 18, 24, 30, and 36 (uniformly distributed within the liver). Apply the mappings of the registrations to manual liver segmentations and compute an overlap measure. An overlap measure reflects degree of alignment in the liver region. We observe that the overlap measures gradually increase from 4 point case and then plateau at 24 points. For the other organs, we use the same number of control points from our previous work as they seem to work reasonably well. Previously we had 36 control points picked by an expert; 17 points in liver, 6 points each for both kidney, and 7 points for spinal cord. We have improved our previous work by systematically choosing control points in the liver, which primarily drives the abdominal registration.

2.3 Scans Used

We have 24 CT scans with available expert manual segmentations. Manual segmentations contain liver, both kidneys, and spinal cord. Since there is only one manual segmentation per scan, we cannot compute inter/intra-observer variability of manual segmentations. A typical CT scan has matrix of $512\times512\times100$ with $1\times1\times3$ mm^3 voxel dimensions. We only recruited CT scans where internal structures (e.g., vessels observed during the arterial and portal venous contrast phases) are visible within the liver. Previously, our CT scans included many scans where no internal structure was visible within the liver. Control points in regions of limited information are meaningless because the gradient of the cost function, i.e., MI, will be very small with respect to deformation resulting from movement of these control points.

2.4 Construction of Atlas

In computing the probabilistic atlas we need to perform many registrations of pairs of CT scans. Registration of two abdominal CT scans is primarily driven by liver as the liver occupies the largest volume in the abdomen. Liver affects the joint histogram and the resulting MI more than other organs. Thus, smaller organs like kidneys are not accurately aligned if simultaneously registered with other organs. By registering each organ separately, better overall registration accuracy is obtained. For these organ-specific sub-registrations, the scans are masked so that the masked scan contains the organ of interest and its immediate vicinity. This masking ensures that organ-specific sub-registrations are driven by information of that organ only. Masked liver is registered using 24 control points, masked kidneys are registered using 6 control points each, and masked spinal cord is registered using 7 control points. After all sub-registration tasks are finished, one final registration is performed between two CT scans using an initial guess obtained from combining the previous four sub-registrations. The initial guess of the final registration step may be inconsistent as sub-registration tasks are only accurate for their respective organs. For example, liver registration step is accurate within liver but it may be erroneous outside the liver, which might affect kidney registrations when all sub-registrations are combined. Thus, we need an extra registration step where all control points are optimized simultaneously. For the final registration step, since we are reasonably close to the intended solution, search range of the optimizer is made smaller than the previous four registration steps. This last registration step is to rectify any inconsistent control point interactions which might have occurred in the four previous sub-registrations. Also note that all registration tasks here are automatic after the user's initial placement of 4 control points.

2.5 Computation of Atlas

Our approach to atlas building is to pick a target and then register all the other scans onto the chosen target. Once all scans are registered onto the target space, we apply the same mappings to the manual segmentations of the CT scans so that all manual segmentations are mapped onto the same target space. The probabilistic atlas is defined on the target space where one measures a probability of an organ occurring in each voxel. For every voxel, we compute the probability from the frequency of occurrence of mapped manual segmentations, e.g. how many times out of the total number of 24 cases the given voxel is liver. We repeat this process for all four organs, thus we have four probability values per voxel. There is significant remaining space in the abdomen CT where a voxel doesn't belong to any of the four organs. For this, we introduce a fifth component, "none of the above". The probability of the fifth component is computed to be 1 minus the sum of the four probability values so that all five probability values add up to one, voxel-wise. Depending on the choice of the target, there may be many atlases from a set of training scans. Previously, we picked a target scan which was thought to best represent the training set by an expert. This time, we introduce a better way to choose the target space based on multidimensional scaling (MDS).

2.6 Multidimensional Scaling and Distance Measure

MDS is a classical statistical tool to produce relative positional locations from a collection of pair-wise distances [29], [30]. The relative locations are accurate up to arbitrary rotate-translate transform. We refer to these references for more information on MDS [31]–[33]. MDS requires pair-wise distances as its input. We have proposed a

distance measure based on bending energy to quantify the distance between two registered scans [10]. The outcome of registration task is a geometric transform. The displacement field is computed by evaluating the geometric transform at every voxel. The geometric distance, hereafter called distance, between two scans is often measured by the roughness of the geometric transform. We have chosen the distance to have invariance to affine transforms. For example, if scans can be registered with an affine transform, then it implies that the scans are essentially composed of the same objects described in different coordinate spaces; thus a value of zero is assigned for the distance between the two. We have chosen bending energy defined by the sum of second partial derivatives of the geometric transform as the distance,

$$d^2 = \sum_{j=1}^{3} \iiint \left(\frac{\partial^2 f_j}{\partial x^2}\right)^2 + \left(\frac{\partial^2 f_j}{\partial y^2}\right) + \left(\frac{\partial^2 f_j}{\partial z^2}\right)^2 dxdydz + 2\sum_{j=1}^{3} \iiint \left(\frac{\partial^2 f_j}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 f_j}{\partial x \partial z}\right) + \left(\frac{\partial^2 f_j}{\partial y \partial z}\right)^2 dxdydz$$

f_1; displacement in x $\quad f_2$; displacement in y

f_3; displacement in z (2)

Second order derivatives ensure invariance to affine transforms. An analytic formula for calculating bending energy is available for TPS [34]. For other geometric transforms, the bending energy may need to be calculated numerically. The defined distance in (2) is strictly not a metric as it doesn't satisfy the first axiom of a metric (i.e., isolation, $d(a, b) = 0$ iff $a = b$) since the distance between two different scans can be zero if two scans are registered by an affine transform. The defined distance satisfies the second axiom (i.e., symmetry, $d(a, b) = d(b, a)$) as switching the order of scans to be registered has no effect on the displacement field in theory. We haven't able to prove or disprove the third axiom (i.e., triangle inequality, $d(a, b) + d(b, c) \geqq d(a, c)$) and leave this to future work. One example of a metric distance is the viscous fluid model where the distance is invariant to only an identity transform [35]. If the requirement of invariance to affine transform is dropped, others have proposed a distance satisfying all three axioms of metric. The distances used in MDS need not be metric, as non-metric distances (e.g., ranking or Riemannian distance) can be used [36]–[38]. Thus, our distance measure can be used in MDS settings.

Given a set of distances in the distance matrix D, where an element of matrix dij refers to the distance between objects i and j, MDS outputs a set of coordinates X in a user specified dimension p that reproduces the distance matrix best in the least square fashion. A standard way of determining the MDS dimension, p, is to perform a sequence of MDS projections, successively increasing the dimension at each iteration, and detecting a knee in the set of fitting errors. This is equivalent to choosing the dimension by thresholding the scree plot of sorted eigenvalues of the distance matrix. With TPS based registration, switching the order of scans in the registration does not yield the inverse transform, and thus it may yield a different distance value, but the discrepancy in distance values is quite small provided that the DOF of TPS is high enough. We take the average value of dij and dji to achieve a symmetric distance matrix.

2.7 MDS Based Target Selection

The ideal target is the one that resides at the mean geometry of the training scans as measured by the bending energy distance. Under this circumstance, the sum of square distances to other scans from the atlas space (i.e., target space) is minimized for the ideal target. Often there may not be a scan at the mean geometry; thus the best approach in picking a target which yields the minimum distance to other scans is to choose the scan that is the closest (i.e., smallest Euclidean distance in the sub-space of MDS output) to the mean geometry. The described approach works only if we know all the relative locations of scans of the training scans so that the location for the mean geometry can be calculated. MDS identifies all the relative locations of the scans from the distance matrix. The mean geometry is computed to be the sample mean of the Euclidean coordinates of the scans. The elements of the distance matrix are determined by the distances of pair-wise registrations. The following is the procedure for N scans,

1. Perform $N(N-1)$ pair-wise, forward and inverse registrations
2. Calculate bending energies from the registrations
3. Form distance matrix D
4. Determine embedding dimension of MDS
5. Apply MDS and find relative locations of scans
6. Calculate mean location of the scans
7. Choose target scan that is the closest to the mean.

Once the best (i.e., the closest to the mean geometry) target is selected, all other scans can be mapped onto the chosen target with ease; this step is trivial since all pair-wise registrations have been computed previously to fill the distance matrix. Pair-wise registration is impacted by the initial placement of control points and the optimizing algorithm. If the initial placement is too far away from the intended solution then it fails to converge to the intended solution. An optimizer with a narrow convergence range has the similar negative effect. Error in the pair-wise registration may lead to possibly not choosing the best target for atlas construction as some of distance measures are erroneous.

3. Results; Atlas Construction

3.1 Variance of Atlas

Our probabilistic atlas is a 5 vector field that resides in the same space of the chosen target. The 5 vector components are the probabilities of liver, right kidney, left kidney, spinal cord, and "none of the above" for each voxel. At a given voxel, all 5 components add up to one. Each vector component measures the probability of an organ presence at that

voxel. If perfect registrations are possible among all training scans, there is no need for an atlas, as information (i.e., manual segmentation) can be carried from one scan to another without error. In that hypothetical situation, the atlas will be the same as the manually segmented dataset with binary probability values of either 0 or 1. In reality, perfect registration is not achievable due to computational and anatomical reasons. Most people have left lobes in their livers while others don't. No one-to-one (i.e., invertible) transform can model such existence/absence of organs. Even if the same structures are present in both scans, it may require many DOF to model the complex shape difference. Most geometric transforms are capable of supporting such a complex transform, but registration still requires adequate underlying grayscale information to support such a high DOF transform. Normally, as the registrations are never perfect, the probabilistic atlas will have intermediate values between 0 and 1. These intermediate values reflect voxels when the organ is present between 0 and 24 occurrences out of a possible 24 cases. The atlas has variance coming from anatomical and computational reasons. Anatomical variance is present due to underlying differences in the population, but computational variance can be reduced if one uses registrations with many DOF. Researchers have observed decrease in atlas variance as one increases DOF of the transform. As the user traverses through the atlas space, the variance of the atlas can be visualized as a rising or falling edge of fractional values between 0 and 1 with variable transition zone. For instance, if one travels from outside the liver into the liver, one would observe a rising edge from 0 to 1 regarding probability of liver presence. The rate of transition reflects the spatial variance of the atlas, i.e. full transition in a short distance implies low variance while a slow transition with respect to distance indicates high variance. An atlas with low variance is preferred to reduce uncertainty.

3.2 Two Atlases; Best and Worst Determined by MDS

Many different atlases may be built from the same set of training scans depending on the choice of target space. With the aid of MDS, we can choose the optimal target. We need to perform $24 \times 23 = 552$ pair-wise registrations to fill up the distance matrix needed for MDS. In essence, we compute all possible pair-wise registration among training scans. Once the distance matrix is ready, MDS is performed with 3 dimensions. The dimension of MDS is determined from sorted eigenvalue plots of the distance matrix. MDS result indicates that the target space closest to the mean geometry is the one labeled "68f3" and the target space furthest from the mean geometry is labeled "1bb1". In Fig. 1, we provide two atlases: one built on space "68f3" (Fig. 1 (a)) and one built on "1bb1" (Fig. 1 (b)). By visual inspection the atlas built on the closest target to the mean, left figure, has less variance then the atlas built on the furthest target, right figure. Notable increased variance can be observed in the right kidney (green hue) and spinal cord (yellow hue) regions. The right kidney in the right figure appears more dispersed than the left figure. Additionally the left lobe of the liver is defined more normally in the atlas on the left using the target "68f3".

(a) (b)

Fig. 1 Two atlases built from the same 24 training CT scans. A mid-hepatic slice out of a 3D volume is shown here. Red corresponds to liver and green, blue, and yellow correspond to right kidney, left kidney, and spinal cord respectively. LEFT; atlas built on target 68f3 (best target), RIGHT; atlas built on target 1bb1 (worst target).

Thus far we have shown a trend that choosing a target space that is the closest to the mean geometry as determined by MDS leads to atlas with low variance. This is still not a validation that MDS based approach chooses the best available target as we have no ground truth regarding which target is the closest to the mean geometry from 24 real CT scans. We have only shown a general trend. However, we propose an alternative validation of MDS based target selection via segmentation performance, which can be measured with respect to manual segmentation. If MDS can find a least biased target, then whatever atlas was built on that target should have the least segmentation error among possible atlases. We will show that the atlas built on the closest target space determined by MDS leads to better segmentation performance.

4. Methods; Segmentation

A segmentation algorithm takes a gray scale input scan and produces an output consisting of a few labels. Here we consider a 5-label model, liver, spinal cord, left kidney, right kidney, and "none of the above", which are the same labels as the atlas. In this paper, we are particularly interested in atlas-based segmentation algorithms where the atlas provides additional side information to the usual grayscale scan. We consider two segmentation algorithms. One is the segmentation by registration method and the other is the maximum a posteriori (MAP) method. Performance of each segmentation algorithm is evaluated with respect to the manual segmentation provided by an expert. We adopt the standard type I and II error measures, false positive rate and false negative rate. A voxel is deemed false positive if the segmentation result indicates presence of an organ while the manual expert segmentation indicates its absence. Likewise, a voxel is deemed false negative if the segmentation result indicates absence of an organ while the manual segmentation indicates its presence. A segmentation algorithm with lower false positive and false negative rates, both preferably closer to zero, is considered better than a segmentation algo-

rithm with higher lower false positive rate and false negative rate. True positive rate is defined as 1 - false negative rate. A pair of values, i.e. the false positive rate and the true positive rate, makes up a single point in a receiver operating characteristic (ROC) curve. The area under the curve (AUC) of the ROC curve is a commonly used measure to compare performances of segmentation algorithms. A segmentation algorithm that produces higher AUC, preferably closer to one, is considered better than a segmentation algorithm with lower AUC.

4.1 Registration Based Segmentation

The first segmentation method we consider is segmentation by registration. This method propagates information in one scan to the other scan via registration. Given a perfect registration between scans, one can carry the manual segmentation information of one scan to the other scan without error. With imperfect registrations, there are errors in the propagated information. Here, we register, i.e. map the target scan into the test scan geometry, and then propagate the atlas information to the test scan. Multiple segmentations are obtained by applying thresholds to the atlas. For each organ, regions where atlas of the organ has probability values above the threshold are taken to be the segmentation of that organ. We apply 5 threshold values, 0.1, 0.3, 0.5, 0.7, and 0.9. Low threshold values yield larger segmented areas and possibly lead to aggressive, over-segmented results. High threshold values yield smaller segmented areas and possibly lead to conservative, under-segmented results. Depending on the threshold, we get 5 different segmentations, which leads to 5 pairs of false positive and false negative rates. These 5 pairs correspond to 5 operating points on a receiver operating curve (ROC), which are then used to compute the AUC. We adopt this simple segmentation method since our purpose is to show the value of an atlas to segmentation. With simple segmentation it is easier to observe the differences in segmentation outcomes depending on the quality of the atlas. Other more sophisticated segmentation algorithms may compensate for a bad quality atlas. Thus, difference in segmentation performance with respect to the quality of the atlas for sophisticated segmentation algorithms may be more difficult to observe.

4.2 MAP Segmentation

The other segmentation method we consider is maximum a posteriori (MAP) segmentation. It is formulated in a Bayesian framework where the atlas information is considered to be the a priori probability of segmented labels. We have proposed this MAP approach combined with Markov Random Field (MRF) for smoothing the segmented outcome [4]. Following is a brief description. Assuming that X is the volumetric label field to be estimated and Y is the volumetric grayscale observations, MAP tries to maximize a probability $\Pr(X|Y)$ which leads to maximizing the probability $\Pr(Yi|Xi)\Pr(Xi)\exp(-\beta\delta i)$ on a voxel-by-voxel basis, where $\Pr(Yi|Xi)$, $\Pr(Xi)$, β, and δi denote voxel-wise conditional probability of grayscale values given a label, voxel-wise atlas prior information, strength of smoothing for MRF, and number of different labels in a 6-voxel neighborhood label field, respectively. Note that we obtain one segmentation outcome per scan if we fix the strength of smoothing β, and thus we will get one, not multiple, operating points on the ROC curve.

Both segmentation methods require the user to register the test scan to be segmented onto the atlas space so that the test scan and the atlas reside in the same space. The registration process is the same one used to construction the atlas, found in Sect. 2.3, so that the atlas information is consistently applied to the test scan.

5. Results; Segmentation

As there are only 24 CT scans with manual segmentation, we adopt a leave-one-out approach when applying segmentation algorithms. We build an atlas with 23 scans and use the atlas information to aid segmentation of the remaining, left out scan. We repeat this process 24 times choosing a different scan as the left out scan. Three different atlases are built for each left out scan to be segmented using the relative positional information provided by the MDS. We build an atlas on the target that is the closest, 13th closest, and furthest from the mean geometry determined by MDS. These three atlases represent the best, intermediate, and the worst atlases possible respectively. MDS is performed with 3 dimensions.

5.1 Sample Segmentation Results

In Fig. 2, we provide sample segmentation results using the best atlas determined by MDS for both segmentation methods. Here the test scan has been registered on to the atlas

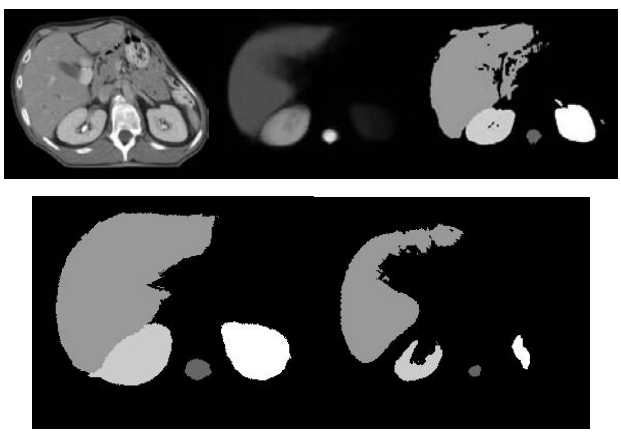

Fig. 2 Mid-hepatic sample slice of segmented output. Note that these results are one slice out of volumetric results. TOP LEFT; CT cross-section mapped into atlas target geometry, TOP MIDDLE; Corresponding atlas cross-section, TOP RIGHT; Segmentation of CT using MAP algorithm, BOTTOM LEFT; Registration-based segmentation using atlas threshold of 10 %, BOTTOM RIGHT; Registration-based segmentation using atlas threshold of 90 %.

space so that segmentation methods can be applied. For the MAP approach, we use strength of smoothing $\beta = 1.5$, 6 voxel neighborhood for MRF, and 0.85 threshold for automatic training. For the registration-based segmentation method, we use threshold values of 0.1 and 0.9. All three segmentation results produce reasonable results. They all eliminate unwanted organs, body wall, bones, intestine and etc., while delineating four organs of interest (i.e., liver, both kidneys, and spinal cord). Comparing two segmented results using registration based approach shows that low threshold value (i.e., 0.1) leads to over segmentation, and high threshold value (i.e., 0.9) leads to under segmentation as predicted.

Performance measures for three segmentation results of Fig. 2 are computed using the manual segmentation as the ground truth and are presented in Table 1.

We also compute the performance measure for other threshold values, 0.3, 0.5, and 0.7, for the registration based segmentation method. These performance measures for both MAP and registration-based methods are plotted using colors in Fig. 3. Registration based segmentation yields 5 operating points on the ROC curve since there are 5 different segmentation results, one for each of the 5 different segmentation thresholds, while the MAP method only yields one operating point (plotted with a round dot) on the curve since there is only one segmentation result. Note that MAP operating point is located at the upper left to the ROC curve of registration based method, which indicates that MAP segmentation generally performs better, i.e. more true positives for the same number of false positives, than the registration-based segmentation. With MAP method, it is hard to obtain a ROC curve as changing the parameters of MAP segmentation yields very small changes in the segmented output and thus small changes in the performance measure. We repeat the whole process of computing performance measures for the sample in Fig. 2 using the atlases built on the 13th closest, i.e. intermediate, target and furthest, i.e. worst, target. Performance measures using the intermediate atlas are plotted in green and performance measures using the worst atlas are plotted in red. Note that the blue plots are located at the upper left to the green plots and the green plots are located at the upper left to the red plots, which implies that segmentation using the best atlas is better than segmentation using the intermediate atlas; additionally segmentation using the intermediate atlas is better than segmentation using the worst atlas. Within using the same atlas the MAP segmentation is better than the registration-based method as the dots (MAP operating point) are located above the operating curve of the registration-based approach). This suggests that additional complexity of MAP approach yields better performance.

Table 1 Segmentation performance as a function of segmentation method.

Segmentation method	Performance measures	
	False positive fraction	False negative fraction
MAP	0.0027	0.0236
Registration to atlas with a threshold of 0.1	0.0001	0.2917
Registration to atlas with a threshold of 0.9	0.0071	0.0113

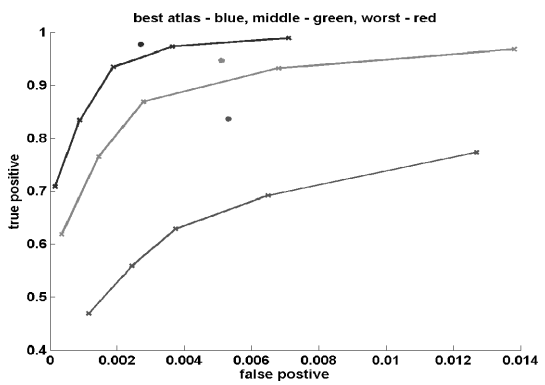

Fig. 3 ROC performance, i.e. true positive fraction vs. false positive fraction, measures of sample segmentation. Blue plots (darkest) are performance measures obtained using the best atlas, green plots (less dark) are performance measures obtained using the intermediate atlas, and the red plots (least dark) are performance measures for using the worst atlas. Registration based segmentation yields 5 operating points on the ROC curve (plotted with "x" and lines) and MAP method only yields one operating point (plotted with a round dot) on the curve.

5.2 Segmentation Results of 24 CT Scans

Thus far we have discussed performance measures derived from one scan using different atlases and different segmentation methods as shown in Fig. 3. Now we move onto applying the two segmentation methods using three different atlases for 24 CT scans using the leave-one-out jack-knifing approach. For registration-based segmentation method, we observe $24 \times 3 = 72$ ROC curves from 24 CT scans and using 3 different atlases. We observe 24 blue, green, and red curves using the best, intermediate, and the worst atlas, respectively, as shown in the top plot of Fig. 4. In the same fashion, we observe $24 \times 3 = 72$ operating points for the MAP approach color coded the same way as shown in the bottom plot of Fig. 4.

5.2.1 Registration-Based Segmentation Results

For the registration-based segmentation approach, AUC is computed using a trapezoid approximation for all 72 ROC curves and then compared. Three groups for comparison are segmentation using the best atlas, intermediate atlas, and the worst atlas, denoted group 1, 2, and 3 respectively. The larger the AUC is, the better the segmentation algorithm is. The AUCs for three groups are reported in Table 2.

The mean AUC of group 1 is larger than the mean of group 2 and the mean of group 2 is larger than mean of group 3. All stdev values are large enough so that the differences between groups are not clear cut. We compare AUCs from using 3 different atlases by standard one-way

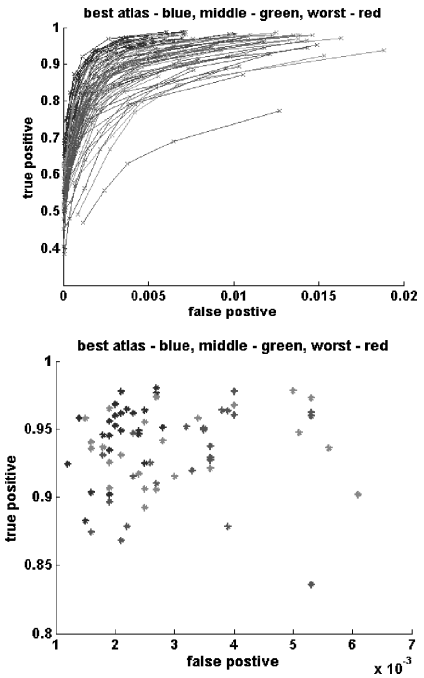

Fig. 4 Performance measures of 24 CT scans. Blue (darkest) plots are performance measures for using the best atlas, green (less dark) plots are performance measures for using intermediate atlas, and red (least dark) plots are performance measures for using the worst atlas. The top figure is the ROC curves for the registration-based segmentation and the bottom figure is for the MAP approach.

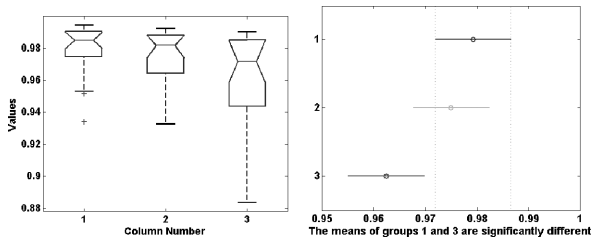

Fig. 5 ANOVA and multiple comparison results of AUCs for registration based segmentation. Left plot is a boxplot where the boxes have lines at the lower quartile, median, and upper quartile values. The whiskers are lines to show the extent of the data. Outliers are marked with red "+". Column 1, 2, and 3 correspond to group 1 (using the best atlas), group 2 (using the intermediate atlas), and group 3 (using the worst atlas) respectively. The right plot is the result of a multiple comparisons analysis. Horizontal bars are estimated intervals for the groups. If the bars overlap they don't have significantly different means. Dots in the center of the bars indicate means of the groups. Y-axis denotes the groups being compared.

Table 2 Performance measures of registration based segmentation.

Target Atlas	mean AUC	stdev AUC
Group 1, best atlas	0.979	0.0159
Group 2, intermediate atlas	0.975	0.0153
Group 3, worst atlas	0.962	0.0294

balanced ANOVA and subsequent multiple comparisons using Tukey-Kramer's "honest differences". The one-way ANOVA shows a p-value of 0.0021, showing that at least one group has its mean significantly different from other groups. Multiple comparisons shows that group 1 is significantly different from group 3. ANOVA and multiple comparisons results are shown in Fig. 5. We observed a similar trend, group 1's mean is significantly different from group 3, using a non-parametric analysis, Kruskal-Wallis method, combined with multiple comparison with a p-value 0.0392. In summary using the best atlas for segmentation (i.e., group 1) is better than using the worst atlas (i.e., group 3). However, using the intermediate atlas (i.e., group 2) doesn't show a significant difference from groups 1 and 3 for the sample size of 24 datasets.

5.2.2 MAP Segmentation Results

For MAP segmentation, AUCs are not available. We report false positive and false negative rates for three groups in Table 3.

The mean false positive rate and false negative rate of

Table 3 Performance Measures of MAP Segmentation.

Atlas Used	False positive rate mean(stdev)	False negative rate mean(stdev)
Group 1, best atlas	0.0021 (0.0004)	0.0526 (0.0247)
Group 2, intermediate atlas	0.0031 (0.0014)	0.0630 (0.0253)
Group 3, worst atlas	0.0033 (0.0011)	0.0744 (0.0370)

group 1 are smaller than the means of group 2 and the means of group 2 are smaller than means of group 3. All standard deviation values are large enough so that the differences among atlas construction groups are not statistically significant. We compare false positive rates and false negative rates for 3 different groups using ANOVA and multiple comparisons in the same fashion as in analyzing AUCs. For false positive rates, ANOVA shows p-value of 0.0003 and multiple comparisons shows that group 1 is significantly different from group 3, as shown in top row plots of Fig. 6. For false negative rates, ANOVA shows p-value of 0.0435 and multiple comparisons also shows that group 1 is significantly different from group 3, as shown in bottom row plots of Fig. 6. In result using the best atlas for segmentation (i.e., group 1) is better than using the worst atlas (i.e., group 3). However, using the intermediate atlas (i.e., group 2) doesn't show a significant difference from groups 1 and 3 for the sample size of 24 datasets.

5.2.3 Comparison of Registration Based and MAP Segmentation

Inspecting the top and bottom plots of Fig. 4 shows that dots in the bottom plot of Fig. 4 reside at the upper left side of the operating points from registration based segmentation in the top plot of Fig. 4. Note the differences in extents of x-axis and y-axis of the top and bottom Fig. 4. This implies that MAP approach has better performance than registration based method in general. However with a simple

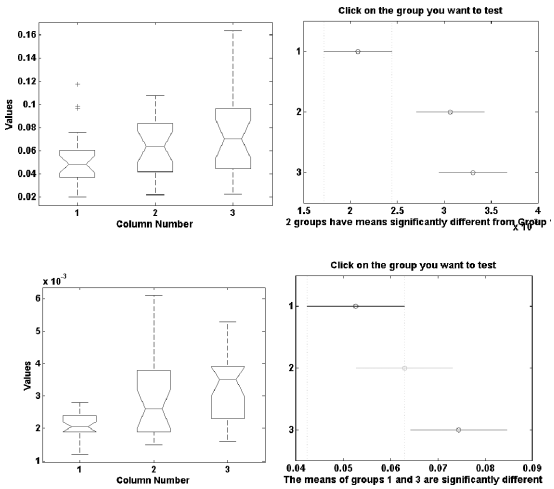

Fig. 6 ANOVA and multiple comparison results of false positive rates and false negative rates for MAP segmentation. Top row shows analysis for false positive rates and bottom row shows analysis for false negative rate. Same figure notations from Fig. 5 apply.

registration-based segmentation, it is easier to observe differences in segmentation outcome depending on the quality of the atlas. Thus, we observe larger range of operating points.

6. Discussion

We built an abdominal probabilistic atlas of 24 CT scans by mapping other training scans onto a chosen target. We choose a target space that is the least biased determined by relative positional information of MDS. MDS requires a distance matrix whose elements are computed from pair-wise registrations. Many pair-wise registration, 24×23 = 552, are needed to fill the distance matrix. Each pair-wise registration took around 4 hours on a Pentium 4 3.0 GHz computer with 4 Gigabytes of memory. Other atlas construction methods [13]–[17] register all training scans at the same time, thus requiring one registration step with huge DOF. It is difficult to compare the computational costs as it is difficult to implement the methods in [13]–[17]. Once the atlas is built, applying the atlas to segment a test scan is less computationally intensive, as the user only needs to register the test scan onto the atlas once. However, if the user wants to add more scans to the training scans, then it requires computing all possible pair-wise registrations with respect to existing training scans.

Our method of choosing a target space via MDS cannot be validated directly as there is no ground truth regarding which target is the closest to the mean geometry from 24 real CT scans. Instead, we provide an alternative validation by segmentation performance. If MDS can find a least biased target, then an atlas built on that target should have the least variance among other atlases constructed using alternative targets. We provide an alternative validation by showing that the atlas built on the closest target determined by MDS leads to better segmentation performance compared to using other atlases built on targets of different MDS distances from the mean. Clearly this performance exists in a continuum where the best and worst atlas targets as judged by MDS are extreme opposites. Although the intermediate atlas seems to show a bias trending worse than the optimal atlas, we are unable to characterize the intermediate atlas as significantly different with a population of only 24 patients.

With both segmentation methods, i.e. registration-based and MAP, there is a significant difference between using the best atlas and the worst atlas. The quantitative usefulness of the atlas to segmentation is determined by the relative positional information of the target in MDS space. While a sophisticated segmentation method like MAP has better performance compared to a simple segmentation method like the registration-based approach, the segmentation performance of both methods is improved by the use of the best atlas. In this paper we demonstrate 1) construction of a probabilistic atlas of abdominal organs using a least biased target space determined by MDS and 2) that the use the best atlas leads to the best segmentation performance, regardless of whether it is done simply by registration with the atlas or MAP segmentation.

References

[1] A. Guimond, G. Subsol, and J. Thirion, "Automatic MRI database exploration and applications," Int. J. Pattern Recog. Artifical Intelligence, vol.11, pp.1345–1365, 1997.
[2] N. Passat, C. Ronse, J. Baruthio, J.P. Armspach, C. Maillot, and C. Jahn, "Semiautomated four-dimensional computed tomography segmentation using deformable models," Med. Phys., vol.32, pp.2254–2261, 2005.
[3] M. Prastawa, J.H. Gilmore, W. Lin, and G. Gerig, "Automatic segmentation of MR images of the developing new born brain," Med. Image Anal., vol.9, pp.457–466, 2005.
[4] H. Park, P.H. Bland, and C.R. Meyer, "Construction of an abdominal probabilistic atlas and its application in segmentation," IEEE Trans. Med. Imaging, vol.22, no.4, pp.483–492, 2003.
[5] E.D. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens, "An information theoretic approach for non-rigid registration using voxel class probabilities," Proc. MICCAI'03, Lect. Notes Comput. Sci., vol.2879, pp.812–820, 2003.
[6] P.M. Thompson and A.W. Toga, "Detection, visualization and animation of abnormal anatomic structure with a deformable probabilistic brain atlas based on random vector field transformations," Med. Image Anal., vol.1, pp.271–294, 1997.
[7] Y. Zhou and J. Bai, "Multiple abdominal organ segmentation: an atlas-based fuzzy connectedness approach," IEEE Trans. Inf. Technol. Biomed., vol.11, no.3, pp.348–52, 2007.
[8] J.L. Boes, P.H. Bland, T.E. Weymouth, L.E. Quint, F.L. Bookstein, and C.R. Meyer, "Generating a normalized geometric liver model using warping," Investigative Radiology, vol.29, pp.281–286, 1994.
[9] J.L. Boes, C.R. Meyer, and T.E. Weymouth, "Liver definition in CT using a population-based shape model," Computer Vision, Virtual Reality and Robotics in Medicine—CVRMed'95, N. Ayache, ed., pp.506–512, Springer-Verlag, Berlin, 1995.
[10] H. Park, P.H. Bland, A.O. Hero, and C. Meyer, "Least biased target selection in probabilistic atlas construction," Proc. MICCAI'05, Lect. Notes Comput. Sci., vol.3750, pp.419–426, 2005.
[11] S. Joshi, B. Davis, M. Jomier, and G. Gerig, "Unbiased diffeomorphic atlas construction for computational anatomy," NeuroImage, vol.23, pp.S151–S160, 2004.
[12] A. Guimond, J. Meunier, and J. Thirion, "Avarage brain models:

a convergence study," Comput. Vis. Image Understand., vol.77, pp.192–210, 2000.
[13] K. Bhatia, J. Hajnal, B. Puri, D. Edwards, and D. Rueckert, "Consistent groupwise non-rigid registration for atlas construction," Proc. IEEE Symposium on Biomedical Imaging, pp.908–911, Arlington, VA, 2004.
[14] E. Learned-Miller, "Data driven image models through continuous joint alignment," IEEE Trans. Pattern Anal. Mach. Intell., vol.28, no.2, pp.236–250, 2006.
[15] C. Studholme and V. Cardenas, "A template free approach to volumetric spatial normalization of brain anatomy," Pattern Recognit. Lett., vol.25, pp.1191–1202, 2004.
[16] C.J. Twining, T. Cootes, S. Marsland, V. Petrovic, R. Schestowitz, and C.J. Taylor, "A unified information—Theoretic approach to groupwise non-rigid registration and model building," Proc. Information Processing in Medical Imaging (IPMI), Lect. Notes Comput. Sci., vol.3565, pp.1–14, 2005.
[17] L. Zollei, E. Learned-Miller, E. Grimson, and W. Wells, III, "Efficient population registration of 3D data," Proc. Computer Vision for Biomedical Image Applications (CVBIA), Lect. Notes Comput. Sci., vol.3765, pp.291–301, 2005.
[18] S. Marsland, C.J. Twining, and C.J. Taylor, "Groupwise non-rigid registration using polyharmonic clamped-plate splines," Lect. Notes Comput. Sci., vol.2879, pp.771–779, 2003.
[19] A. Tarachandani and D. Boltz, "Review of the basic image processing and segmentation techniques for biological images," J. Imaging Science and Technology, vol.50, pp.233–242, 2006.
[20] S.C. Zhou and A. Yuille, "Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol.18, no.9, pp.884–900, 1996.
[21] J. Marroquin, B. Vemuri, S. Botello, F. Calderon, and A. Fernandez-Bouzas, "An accurate and efficient bayesian method for automatic segmentation of brain MRI," IEEE Trans. Med. Imaging, vol.21, no.8, pp.934–945, 2002.
[22] L. Christopher, E. Delp, C. Meyer, and P. Carson, "3D Bayesian ultrasound breast image segmentation using the EM/MPM algorithm," Proc. IEEE International Symp on Biomed Imag, pp.86–98, 2002.
[23] J.L. Boes, T.E. Weymouth, and C.R. Meyer, "Multiple organ definition in CT using a Bayesian approach for 3D model fitting," in SPIE Vision Geometry IV, pp.244–251, 1995.
[24] D.L.G. Hill, P.G. Batchelor, M. Holden, and D.J. Hawkes, "Medical image registration," Physics in Medicine and Biology, vol.46, pp.R1–R45, 2001.
[25] F.L. Bookstein, "Principal Warps: Thin-plate splines and the decomposition of deformations," IEEE Trans. Pattern Anal. Mach. Intell., vol.11, no.6, pp.567–585, 1989.
[26] C.R. Meyer, J.L. Boes, B, Kim, P.H. Bland, K.R. Zasadny, and P.V. Kison et al., "Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin-plate spline warped geometric deformations," Med. Image Anal., vol.1, no.3, pp.195–206, 1997.
[27] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, Numerical Recipes in C: The Art of Scientific Computing, pp.305–337, Cambridge University Press, Cambridge, 1988.
[28] D. Rueckert, L.I. Sonoda, C. Hayes, D.L.G. Hill, M.O. Leach, and D.J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," IEEE Trans. Med. Imaging, vol.18, no.8, pp.712–721, 1999.
[29] Matlab: Classical Multidimensional Scaling (CMDSCALE). Mathworks: Natick, MA 01760-2098, 2007.
[30] W.S. Torgerson, "Multidimensional scaling: I. Theory and method," Psychometrika, vol.17, pp.401–409, 1952.
[31] F.W. Young and R.M. Hamer, "Theory and application of multidimensional scaling," Eribaum Associates, 1994.
[32] J.K. Lanctot, "Using multi-dimensional scaling to improve machine learning performance," Abstracts of Papers of the American Chemical Society vol.225, pp.U782–U783, 2003.
[33] S.P. Borgatti, http://www.analytictech.com/borgatti/mds.htm, 1997.
[34] F.L. Bookstein, "Quadratic variation of deformations," Inf. Process. Medical Imaging, pp.15–28, 1997.
[35] G.E. Christensen, R.D. Rabbitt, and M.I. Miller, "Deformable templates using large deformation kinetics," IEEE Image Proc., vol.5, no.10, pp.1435–1447, 1996.
[36] X. Pennec, "Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements," J. Mathematical Imaging and Vision, vol.25, pp.127–154, 2006.
[37] X. Pennec, R. Stefanescu, V. Arsigny, P. Fillard, and N. Ayache, "Riemannian elasticity: A statistical regularization framework for non-linear registration," Medical Image Computing and Computer-Assisted Intervention, vol.3750, pp.943–950, 2005.
[38] J.D. Carroll and P. Arabie P, "Multidimensional scaling," Annual Review of Psychology, vol.31, pp.607–649, 1980.

Hyunjin Park has received his B.S. in electrical engineering (1997) from Seoul National University, Seoul, Korea, M.S. in biomedical engineering (2000), and Ph.D. in biomedical engineering (2003) from University of Michigan, Ann Arbor, MI, USA. He was a research faculty with University of Michigan, Radiology between 2004 and 2009. He is currently an assistant prof. with Dept. of Biomedical Eng., Gachon Univ. of Medicine and Science, Incheon, Korea.

Alfred Hero has received his B.S. (1980) from Boston University and Ph.D. (1984) from Princeton University, both in Electrical Engineering. Since 1984 he has been with the University of Michigan, Ann Arbor, where he is a Professor in the Department of Electrical Engineering and Computer Science. He is a Fellow of the Institute of Electrical and Electronics Engineers.

Peyton Bland has received his B.S. in engineering science (1971) from Trinity University, San Antonio, Texas, M.S. in bioengineering (1976), and Ph.D. in bioengineering (1981) from University of Michigan, Ann Arbor, USA. He has been a research faculty in Dept of Radiology at University of Michigan since 1984.

Marc Kessler has received his Ph.D. from University of California, Berkeley (1989). He is currently is an associate professor at Dept of Radiation Oncology, University of Michigan.

Jongbum Seo has received his B.S. in electrical engineering (1999) from Seoul National University, Seoul, Korea, M.S. in biomedical engineering (2001), and Ph.D. in biomedical engineering (2004) from University of Michigan, Ann Arbor, MI, USA. He is currently an assistant prof. with Dept. of Biomedical Eng., Yonsei University, Wonju, Korea.

Charles Meyer has received his B.S. in electrical engineering (1967), M.S. in electrical engineering (1969) from Colorado State University, Ft. Collins, Colorado, and Ph.D. in biomedical engineering (1972) from Iowa State University, Ames, Iowa, USA. Since 1981 he has been with Dept of Radiology at University of Michigan, where he is a professor.