Hierarchical Scale-Based Multiobject Recognition of 3-D Anatomical Structures

Ulas Bagci, *Member, IEEE*, Xinjian Chen, and Jayaram K. Udupa*, *Fellow, IEEE*

***Abstract*—Segmentation of anatomical structures from medical images is a challenging problem, which depends on the accurate recognition (localization) of anatomical structures prior to delineation. This study generalizes anatomy segmentation problem via attacking two major challenges: 1) automatically locating anatomical structures without doing search or optimization, and 2) automatically delineating the anatomical structures based on the located model assembly. For 1), we propose intensity weighted ball-scale object extraction concept to build a hierarchical transfer function from image space to object (shape) space such that anatomical structures in 3-D medical images can be recognized without the need to perform search or optimization. For 2), we integrate the graph-cut (GC) segmentation algorithm with prior shape model. This integrated segmentation framework is evaluated on clinical 3-D images consisting of a set of 20 abdominal CT scans. In addition, we use a set of 11 foot MR images to test the generalizability of our method to the different imaging modalities as well as robustness and accuracy of the proposed methodology. Since MR image intensities do not possess a tissue specific numeric meaning, we also explore the effects of intensity nonstandardness on anatomical object recognition. Experimental results indicate that: 1) effective recognition can make the delineation more accurate; 2) incorporating a large number of anatomical structures via a model assembly in the shape model improves the recognition and delineation accuracy dramatically; 3) ball-scale yields useful information about the relationship between the objects and the image; 4) intensity variation among scenes in an ensemble degrades object recognition performance.**

***Index Terms*—Active shape model, graph-cut, image segmentation, intensity standardization, local scale, object recognition, principal component analysis, three-dimensional (3-D) shape models.**

I. Introduction

THE AIM in *model based segmentation* is to build a model which contains information about the expected shape or appearance of the anatomical structure of interest and match the model to new images. Model based techniques can dramatically improve the efficiency of the recognition and quantitative analysis of anatomical structures compared to manual methods.

Manuscript received November 04, 2011; accepted December 10, 2011. Date of publication December 23, 2011; date of current version March 02, 2012. The work of J. K. Udupa was supported by the National Institutes of Health under Grant HL105212. *Asterisk indicates corresponding author.*

U. Bagci is with the Center for Infectious Disease Imaging, Department of Radiology and Imaging Sciences, National Institutes of Health, Bethesda, MD 20892 USA.

X. Chen is with the Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242 US.

*J. K. Udupa is with Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104 US.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2011.2180920

The segmentation process as a whole can be thought of as consisting of two tasks: *recognition* and *delineation*. Recognition is the process of determining roughly "where" the object is and to distinguish it from other object-like entities in the image [19]. Although delineation—the act of defining the spatial extent of the object region/boundary in the image—is the final step, an efficient recognition strategy is a key for successful delineation. In this paper, the problem of anatomical object recognition (or *anatomy recognition* in short) is tackled through the identification of pose (i.e., orientation, scale, and position) of objects automatically in a hierarchical platform. The proposed recognition method is named hierarchical ball-scale based multiobject recognition (HSMOR). In this paper, we summarize our contributions in two phases. In the first phase, we define HSMOR framework by combining three approaches: first, using coarse to fine recognition strategies to build an efficient model based recognition algorithm; second, incorporating a large number of anatomical structures into the recognition algorithm to yield quick, robust, and accurate segmentations, and third, using scale information to build reliable relationship information between shape and texture patterns that facilitates accurate recognition of single and multiple objects without using optimization methods. In the second phase, we analyze the generalizability of the proposed recognition method for different imaging modalities and identifying modality specific difficulties in anatomical structure recognition process.

The rest of the paper is organized as follows. Section II reviews the related studies in the literature and an overview of our approach. Section III describes the shape model. This is followed by a description of the theoretical fundamentals of our approach including the relationship between shape and intensity structure systems in Section IV. We present the experimental results for recognition experiments and discussion in Section V. In Section VI, we explore the effect of intensity nonstandardness on recognition of anatomical structures, which is followed by a conclusion in Section VII.

II. Related Works and Overview of the Proposed Approach

A. Related Works

Some model based segmentation methods rely on initial placements of the models in the image by experts [1], [2], where user interaction guides the placement process by roughly aligning the position and orientation of the model with the data. However, user interaction often falls short for many segmentation algorithms and a more specific localization is usually required. Similarly, the "Graph-Cut" and "Fuzzy Connectedness" approaches [19], [31], [30], [35], [36] offer manual recognition, in which foreground and background or objects are

specified through user-interactions. User-placed seed-points offer a good recognition accuracy especially in the 2-D case; however, the main drawback of these approaches is that the segmentation results can be unpredictable along weak edges, and the delineation may "leak" into non-object territories. The object of interest is not known geographically by these methods, and the user action specifies only roughly the location of the centres of the objects but neither their orientation, scale, nor geographical layout.

As an alternative to the manual methods, model based methods can be employed for initialization/recognition. The goal in model based recognition is to effectively locate the previously built model in any given image. In recent years, a number of methods have been developed to tackle this problem in efficient ways. For example, in [3], the position of an organ model (i.e., liver) is estimated by its histogram. In [4], the generalized Hough transform is successfully extended to incorporate variability of shape for a 2-D segmentation problem. Although attempting to translate anatomical information into the segmentation framework is promising, these approaches have many drawbacks such as converging to a local minimum during optimization, large search space, high computational cost, and infeasible platform for multiobject segmentation. Two other approaches are the widely known active shape model (ASM) andactive appearance model (AAM) [5], [6]. In ASM, after a statistical model of shape variation is built, a number of hypotheses are made to give approximate locations of the model points. The major drawback of the models is that non-object areas are not taken into account in these models to provide a context for objects.

Atlas based methods are also used to define an initial position of the model. In [7], affine registration is performed to align the data into an atlas to determine the initial position for a shape model of a knee cartilage. In [8] and [9], an image based anatomical atlas (model image) is described such that the model image deforms to fit new images by minimizing intensity differences between voxels. However, an elastic deformation cost is needed to regularize the problem. More recently, probabilistic models such as regression forests [10] and marginal space learning [11] based methods have received interest due to their computational efficiency in detecting and locating organs. However, all these methods are based on exhaustive search and optimization of the constructed models. Due to the large search space and numerous local minima, conducting a global search on the entire image often becomes not feasible. Furthermore, all the methods above are modality specific, hence different strategies for feature extraction pertaining to the imaging modality, and global search methods may be necessary. For instance, MRI has unique challenges such as noise, inhomogeneity, and nonstandardness, however, CT does not have inhomogeneity and nonstandardness issues. Therefore, a general, robust, efficient, and fully automatic recognition strategy for 3-D objects remains a challenging goal. To the best of our knowledge, the presented work is the only existing study for 3-D images attempting to locate objects of interest in a given image without any search or optimization.

B. Overview of Approach

The proposed anatomy recognition framework consists of three phases: training, coarse recognition, and fine recognition.

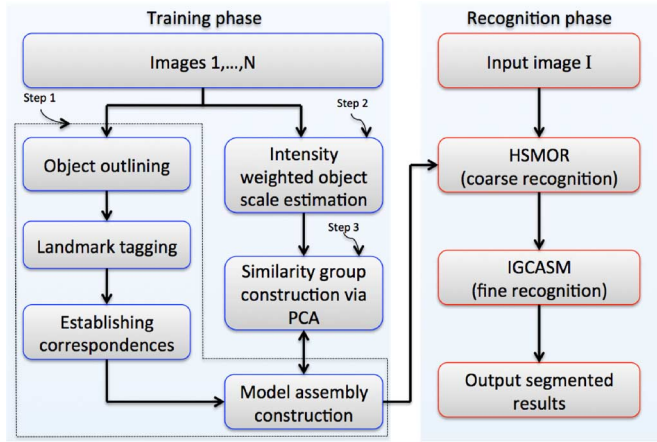

Fig. 1. Flowchart of the proposed multiobject segmentation system.

Fig. 1 shows all three phases and their interactions schematically. The training phase includes three steps: A shape model [model assembly (MA)] is constructed through modelling the shape information of anatomical structures in the first step. In the second step, a similarity group between shape and appearance of anatomical objects is built by extracting hierarchical geometric patterns from grey level images and encoding their appearance through a ball-scale (b-scale) based object encoding method. In the third step, for each shape and appearance pattern set, a relationship function is constructed based on the proposed similarity group. Relationship functions for each shape and appearance pattern set in the training set are used to estimate the mean relationship and are used to determine the location of actual shape patterns for any given test image. The first step of the proposed HSMOR method is explained in Section III, and the second and third steps are explained in Section IV in detail.

In the coarse recognition phase, we roughly localize the MA through using the mean relationship function of the similarity group constructed in the training phase. Finally, the object shape information generated from the training phase and the pose vector of the MA generated from the coarse recognition phase are integrated into the delineation platform where an iterative graph-cut active shape model (IGCASM) algorithm is used for refining the recognition. This step may be called either fine recognition or delineation. The details of each phase are given in the following sections.

III. HSMOR: Shape Modelling

Since model-based recognition of anatomical structures needs incorporation of a prior knowledge, a statistical shape model of anatomical structures [5] (i.e., ASM) is constructed and integrated into the segmentation framework. As seen from the left column of training phase in Fig. 1, there are four parts in constructing MA: A) object outlining, B) landmark tagging, C) establishing landmark and slice correspondences, and D) model assembly construction.

A. Object Outlining

Following Kendal [12], we extracted the shapes of objects through manual outlining by expert radiologists using the Live-Wire algorithm [13], and all information about location, size,

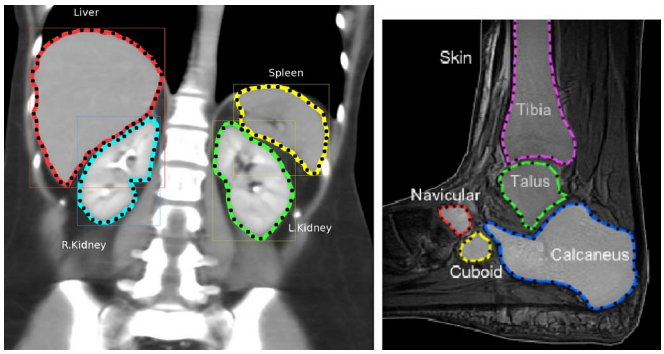

Fig. 2. A CT slice of the abdominal region with landmark-tagged organs (liver, spleen, and left and right kidney) is shown on the left. An MR slice of the foot with landmark-tagged bones (calcaneus, tibia, cuboid, navicular, and talus is shown on the right.

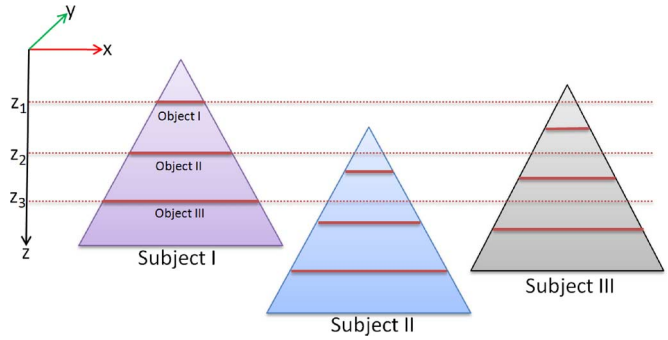

Fig. 3. Slice location (z) for a particular object may correspond to anatomically different slices in different subjects. The positioned slices ($z1$, $z2$, $z3$) of three different objects in the first subject do not correspond to the same locations for the other two subjects, where actual corresponding 2-D slices are indicated by thick lines.

and rotation (*the pose of the shape*) has been filtered out so that we ensure that the variability is from shape changes only and not due to pose differences. This is achieved by aligning all training objects to a common position, orientation, and scale using an appropriate registration technique. A common alignment technique used is an affine transformation using kappa statistics [14]. The affine transformation A consists of seven parameters: three for translation, three for orientation, and one for scaling. Note that only one parameter is used for scaling to represent the relative size of the objects. The main reason for this use is the fact that, if more general affine transformations are used (such as 9 and 12 parameter transformations involving independent scaling and shear in different directions), then the shapes we wish to model may be compromised. That is, it is not guaranteed that the intrinsic structure of the shape is preserved if isotropic scaling is not used. Furthermore, it has been shown in [44] that it is easier to establish correspondences between two shapes that are isotropic than between two shapes with different anistropic scales. When two shapes have different anisotropic scales, it is harder to establish correct correspondences between the two, therefore, matching and localization methods that depend on correspondences for evaluating model similarity will be inaccurate in that case [44].

B. Landmark Tagging

The statistical modelling of shape requires a common description of geometry of the different shapes. This is handled by marking the location of homologous features in each shape. This process is called *landmark tagging* or *landmarking* for short [15]. Although we chose the *landmarking method* to represent shape data due to its *simplicity*, *generality*, and *efficiency*, other shape representation strategies such as meshes [39], medial representations (m-reps) [1], spherical harmonics (SPHARM) [40], and nonuniform rational B-splines (NURBS) [41] can be used as well to represent the shapes in constructing statistical shape models. Fig. 2 shows annotated landmarks for four different organs (liver, right kidney, left kidney, spleen) in a CT slice of the abdominal region, and five different bones in an MRI slice of the foot. The number of landmarks for an object is determined based on its size; for example, more for liver than kidney.

C. Establishing Correspondences

A good statistical shape model requires a precise calculation of shape statistics over the training set. Hence, landmark correspondence must be established among the samples in the training set. Although landmark correspondence (homology) is usually established manually by experts, it is time-consuming, prone to errors, and often restricted to 2-D objects [16], [17], [12]. Because of these limitations, a semi-automatic landmark tagging method, *equally-spaced landmark tagging*, is used to establish correspondence among landmarks in our experiments [17]. Although this method is proposed for 2-D objects, and equally spacing a fixed number of points for 3-D objects is much more difficult, we use this technique in a pseudo-3-D manner, where the 3-D object is annotated slice-by-slice. Equally-spaced landmark tagging on training shape boundaries (contours) starts with selecting an initial point on each shape sample and equally spacing a fixed number of points on each boundary automatically [16]. Each landmark must be located by following the same guidelines for all the example shapes of the training set. Selecting the starting landmark has been done manually by annotating the *same anatomical point* (for example: the left-most corner of the left kidneys, the bottom corner of the spleens, etc.) for each shape in the training set. In place of the above method, any fully automated method may also be used such as SPHARM, NURBS, m-reps, etc.

Yet, another correspondence issue appears in the pseudo-3-D method: the same physical location of slices in one object does not necessarily correspond to the same physical location in another object of the same class (see Fig. 3). Not only large anatomical variability from subject to subject but also the position of the objects within the body (e.g., great variability in the location of the kidneys in the body) makes the selection of anatomically corresponding slices difficult. In order to provide anatomical correspondence among 2-D slices of 3-D objects, a careful selection procedure was devised for use by an expert in the training step [18]. This is a much simpler 1D correspondence problem which is easier and simpler to tackle than even the 2-D point correspondence problem.

D. Single and Multiobject 3-D Statistical Shape Models

In ASM, the characteristic pattern of a shape class is described by the average shape vector (mean shape) and a linear combination of eigenvectors of the covariance matrix of the shape vectors around the mean shape. In multiple-object ASM (MA), each model $\mathbb{M}_i$ for the ith object class can be parametrized with a mean shape $\overline{\mathbf{x}}_\mathbf{i}$ and the covariance matrix Λ_i as $\mathbb{M}_i = (\overline{\mathbf{x}}_\mathbf{i}, \Lambda_i)$ [5]. Each object class brings its unique ASM model into the framework. Therefore, MA can be expressed as a set of models of the form: MA $= \{\mathbb{M}_1, \ldots, \mathbb{M}_M\}$, where M denotes the number of objects considered in the model assembly and each model $\mathbb{M}_i$ consists of a mean shape $\overline{\mathbf{x}}_\mathbf{i}$ and allowable variations given by the covariance matrix Λ_i for object O_i, $1 \leq i \leq M$.

In the training part, we select the objects O_i such that $(O_i \cap O_j) = \emptyset, 1 \leq i \neq j \leq M$. Note that

$$\begin{aligned}(O_i \cap O_j) = \emptyset &\Leftrightarrow (A(O_i) \cap A(O_j)) \\ &= \emptyset \Leftrightarrow (A(IN(\mathbf{x_i})) \cap A(IN(\mathbf{x_j}))) = \emptyset\end{aligned}$$

where $A(.)$ denotes the affine transformation and $IN(\mathbf{x_i})$ denotes the interior of the object defined by shape $\mathbf{x_i}$. Since objects are not aligned separately, their spatial relations before and after alignment do not change. This fact leads to $IN(\overline{\mathbf{x}}_\mathbf{i}) \cap IN(\overline{\mathbf{x}}_\mathbf{j}) = A(IN(\overline{\mathbf{x}}_\mathbf{i})) \cap A(IN(\overline{\mathbf{x}}_\mathbf{j}))) = \emptyset$.

IV. HSMOR: Relationship Between Shape and Intensity Structure System

The HSMOR method allows us to extract hierarchical geometric patterns from grey level images by encoding their appearance and relate this information with true geometric (shape) patterns. The method is based on a similarity group between shape and appearance in the same configuration space, which examines the similarity of regular structures in shape and appearance; therefore, these are called *shape and intensity structure systems*, respectively. Since we represent true and extracted geometric patterns by structured forms that capture much of the salient information of the patterns, there is no need to have exhaustive search algorithms. Hence, the proposed HSMOR method is extremely efficient in providing quick placement of the model for any given image. Patterns from shape and appearance can then be related by independently computing each of their structural systems. For each shape and appearance pattern set, a relationship is defined based on the proposed similarity group. The relationship functions are used to obtain the mean relationship and are used to estimate the pose of true geometric patterns in any given test image. Since extracted geometric patterns are elements of a pattern family which can be thought of as *images modulo the variances* represented by the similarity group proposed, they can naturally be considered as desirable image features to roughly identify the relationship of patterns in terms of scale, position, and orientation. Thus, we conjecture that creating a pattern family that includes rough object information together with region information yields coarse bases for the recognition of objects. For this purpose, observing the grey level images without doing explicit segmentation, a rough but definitive representation of objects, is possible by local scale-based approaches [20].

A. Intensity Weighted Ball Scale Encoding With a Down-Sampling Approach

We integrate locally adaptive scale information of object regions into the recognition process to produce geometric patterns. Based on continuity of homogeneous regions, we roughly identify geometric properties of objects, namely scale information, and represent the actual images with this new representation, called scale images, e.g., ball-scale [20], tensor-scale [25], generalized-scale images [33]. After scale based filtering, resultant rough objects can be used as prior shape information to be integrated into the whole segmentation process because scale images identify structures embodied in the images roughly.

Among local scale based approaches, the b-scale is the simplest form, and has been shown to be useful in image segmentation [19], filtering [20], inhomogeneity correction, and image registration [21]. The main idea in b-scale encoding is to determine the size of local structures at every voxel in an image as the radius of the largest ball centered at the voxel within which intensities are homogeneous under a prepecified region-homogeneity criterion. Inspired from this idea, we incorporate appearance information into this rough knowledge explicitly to characterize scale information of local structures. The proposed method is called *intensity weighted b-scale* or *wb-scale* for short. With this modification, wb-scale filtering allows us to distinguish objects of the same size by their appearance information. As a result, object scale information is enriched with local intensity values.

Assume that we represent a scene as $\mathcal{C} = (C, f)$ where C is a 3-D rectangular array of voxels and f is a function that assigns to every voxel an image intensity value. The homogeneity between two nearby voxels c and d in a scene $\mathcal{C}$ can be characterized by $|f(c) - f(d)|$ [33] or as some monotonically non-increasing function (W_ψ) of $|f(c) - f(d)|$. Several functional forms can be used for (W_ψ) including step functions, normalized or unnormalized Gaussian functions, etc. In this study, we used a zero-mean, unnormalized Gaussian function with a standard deviation of σ_ψ. A hyperball $B_{k,\nu}(c)$ of radius $k \geq 0$ with center at $c \in C$ in a scene $\mathcal{C}$ is defined by

$$B_{k,\nu}(c) = \left\{ e \in C \middle| \sqrt{\sum_{i=1}^{n} \frac{\nu_i^2 (c_i - e_i)^2}{\min_j [\nu_j^2]}} \leq k \right\}. \quad (1)$$

For a hyperball $B_{k,\nu}(c)$ defined above (of any radius $k > 0$ and centered at c), we define a fraction $FO_{k,\nu}(c)$ ("fraction of object"), indicating the fraction of the ball boundary occupied by a region which is sufficiently homogeneous with the voxel c, by

$$FO_{k,\nu}(c) = \frac{\sum_{e \in B_{k,\nu}(c) - B_{k-1,\nu}(c)} W_\psi(|f(c) - f(e)|)}{|B_{k,\nu}(c) - B_{k-1,\nu}(c)|} \quad (2)$$

where $\nu = (\nu_1, \nu_2, \nu_3)$ indicates the size of the voxel, and $|B_{k,\nu}(c) - B_{k-1,\nu}(c)|$ is the number of voxels in $B_{k,\nu}(c) -$

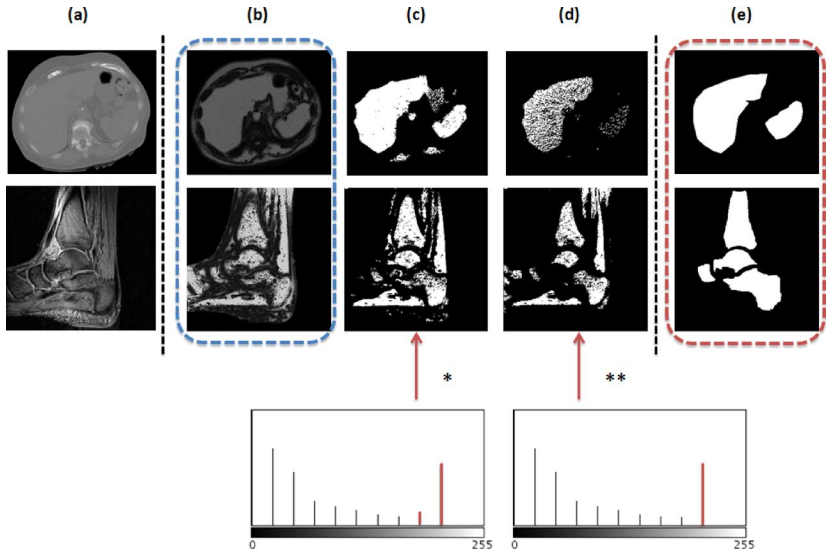

Fig. 4. (a) Original grey-level images: scenes of an abdominal CT image (first row) and a foot MR image (second row). (b) Corresponding wb-scale scenes. (c,d) Thresholded wb-scale images based on the selected object scales (red arrows in histograms). (e) Ground truth delineation of anatomical structures corresponding to CT and MR images shown in (a). (a) original images. (b) wb-scale images. (c) thresholded* wb-scale images. (d) thresholded** wb-scale images. (e) ground truth.

$B_{k-1,\nu}(c)$. The algorithm for wb-scale estimation is presented below.

Algorithm Intensity Weighted Object Scale Estimation (IWOSE) **[18]**

Input: $c \in C$ in a scene $\mathcal{C} = (C, f)$, W_ψ, a fixed threshold $thrs$

Output: wb-scale value: $r'(c)$, b-scale value: $r(c)$

1: **Begin**

2: Set $k = 1$

3: **While** $FO_{k,\nu}(c) \geq thrs$ **do**

4: Set k to $k + 1$

5: **EndWhile**

6: Set $r(c)$ to k

7: Output $r'(c) = f(c)r(c)$

8: **End**

where $r(c)$ and $r'(c)$ indicate b-scale and wb-scale value of the voxel c. A detailed description of the characteristics of homogeneity function W_ψ and $FO_{k,\nu}$ are presented in [20]. In all experiments, we use a zero-mean unnormalized Gaussian function for W_ψ. Following the recommendation in [19], $thrs = 0.85$ is chosen. To reduce computation, we use a multilevel platform where only down-sampled grey level images are used to create wb-scale images. Therefore, the proposed local structure estimation method is called *wb-scale encoding with a down-sampling approach*. The sensitivity of this process is examined in the experimental results section.

B. Positioning Shape Within Image Intensity Structures

The intensity weighted b-scale images, $\mathcal{C}_{wb} = (C, f_{wb})$, can be considered to denote "(intensity weighted) rough objects" because b-scale encoding defines objects roughly and provides object scale estimation based on the continuity of intensity homogeneity. Although this estimation is rough, we hypothesize that there is an explicit relation between this coarse information and the actual object definition (i.e., fine information) in the image. Note that "fine object" is the truly delineated object itself and the process of coarse-to-fine object extraction is equivalent to the whole segmentation process.

The relationship function acts as an equivalence relation of similarity between thresholded $\mathcal{C}_{wb}$ scenes and true shape patterns (ground-truth). Although the relationship can be built and evaluated at any object scale, the selection of higher values of wb-scale or b-scale values ($r'(.)$ or $r(.)$) yields patterns from large scale objects, and the patterns from small scale objects are eliminated. This is desirable because the patterns emerging from large objects are more reliable in terms of identifying scale, location, and orientation of the objects for recognition. Experiments on different selection procedures based on $r'(.)$ or $r(.)$ support the reliability of these patterns due to their global regularity property, as shown in Fig. 4. Note that the histogram of the b-scale image contains only the information about the radius of the balls, therefore, it is fairly easy to eliminate small ball regions and obtain a few largest balls by applying simple *thresholding* to the b-scale or intensity weighted b-scale scene (see right column of Fig. 4). Particularly in this case, thresholding can be used effectively to retain reliable object information. The patterns pertaining to the largest balls retained after thresholding have strong correlations with the truly delineated objects shown in the last rows of the figure.

In recognition, as the aim is to recognize "roughly" the whereabouts of an object of interest in the scene, and also since the trade-off between locality and conciseness of shape variability will be modulated in the delineation step, it will be sufficient

to use concise bases produced by principal component analysis (PCA) without considering localized variability of the shapes. For the former case, on the other hand, it is certain that analyzing variations for each subject separately instead of analyzing variations over averaged ensembles leads to exact solutions where specific information present in the particular image is not lost.

1) Relationship Function: In order to find the translation, scale, and orientation that best align the shape structure system of the model with the intensity structure system of a given image, we learn the similarity of shape and intensity structure systems in the training images via PCA to keep track of translation and orientation differences. We use the bounding box approach to find scale similarity. In the bounding box approach, the *real physical size* of the segmented objects and the structures derived from thresholded intensity weighted b-scale (twb-scale) images are used. For orientation analysis, parameters of variations are computed via PCA. The principal axes (PA) systems of the shape and intensity structures, denoted $\mathbf{PA}_{\text{shape}}$ and $\mathbf{PA}_{\text{intensity}}$, respectively, have an origin and three axes representing the inertia axes of the structure. For the PA systems of the same subject, the relationship function $\mathbf{F}$ that maps $\mathbf{PA}_{\text{intensity}}$ into $\mathbf{PA}_{\text{shape}}$ can be decomposed into the form $\mathbf{F} = (\mathbf{s}, \mathbf{t}, \mathbf{R})$, where $\mathbf{t} : (t_x, t_y, t_z)$ is the translation component, $\mathbf{s}$ is a scale component, and $\mathbf{R} : (R_x, R_y, R_z)$ represents three rotations. We observe that $\mathbf{F}$ can be split into three component functions $\mathbf{f_1}, \mathbf{f_2}, \mathbf{f_3}$, corresponding to scale, translation and rotation, respectively. In the following subsections, how the scale, translation, and orientation components of $\mathbf{F}$ are learnt is explained in detail.

Estimation of the Scale Function: The *bounding Box* enclosing the objects of interest for each subject in the training set is used to estimate the real physical size of the objects in question [26]. The length of the diagonal is used for estimating the scale parameter. The mean scale parameter $\overline{s}$ and standard deviation of scale parameter $std(s)$ are used to obtain an interval for the estimation.

Estimation of the Translation Function: This is solely based on forming a linear relationship between the centroids of the objects of interest obtained from the manually segmented images in the training set and the thresholded wb-scale images. These centroids are denoted by $\mathbf{c}^{\mathbf{i}}_{\text{shape}}$ and $\mathbf{c}^{\mathbf{i}}_{\text{intensity}}$, respectively. By averaging the translational vector over N subjects in the training set, we get the mean translation vector as

$$\overline{\mathbf{t}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{c}^{\mathbf{i}}_{\text{shape}} - \mathbf{c}^{\mathbf{i}}_{\text{intensity}}). \tag{3}$$

For any given test image $\mathcal{C}$, we estimate the centroid of objects in it by $\mathbf{c}'_{\text{shape}} = \mathbf{c}_{\text{intensity}} + \overline{\mathbf{t}}$, where $\mathbf{c}_{\text{intensity}}$ can be determined from the thresholded wb-scale image of $\mathcal{C}$. We assume that the real centroid $\mathbf{c}_{\text{shape}}$ should lie in the region $\mathbf{c}'_{\text{shape}} \pm std(\mathbf{t})$. We use appearance based centroids to build the $\mathbf{f_2}$ component of $\mathbf{F}$ in our experimental set-up to increase the correlation of two structures by considering not only shape features, but also texture features .

Estimation of the Orientation Function: Since the shape and intensity structure systems constitute an orthonormal basis, and assuming that the translation between the two systems is eliminated by using $(\mathbf{c}^{\mathbf{i}}_{\text{shape}} - \mathbf{c}^{\mathbf{i}}_{\text{intensity}})$ for each image i in the training set, the two systems are related by

$$\mathbf{PA}_{\text{shape}} = (\mathbf{R})(\mathbf{PA}_{\text{intensity}}) \tag{4}$$

where $\mathbf{R}$ is an orthonormal rotation matrix carrying information about the relative positions of shape and intensity structure systems in terms of their Euler angles.

A set of N segmented training images and their corresponding intensity weighted b-scale images are used to find their PA systems so that we can relate them by computing the orthogonal rotation matrices $\mathbf{R_i}$ that relate $\mathbf{PA}_{\mathbf{shape_i}}$ to $\mathbf{PA}_{\mathbf{intensity_i}}$ for $i = 1, \ldots, N$. To obtain the basic population statistics over these N subjects, we need to compute the mean and standard deviation of the N rotation matrices $\mathbf{R_i}$, $i = 1, \ldots, N$. Since three-dimensional orientation data are elements of the group of rotations that generally are given as a sequence of unit quaternions, or as a sequence of Euler angles, etc., the group of rotations does not form a Euclidean space, but rather a differentiable manifold. In our case, in analogy with the mean in Euclidean space, mean rotation is defined to be the minimizer of the sum of the squared geodesic distances from the given rotations in the spherical space. The mean rotation $\mathbf{R}^*$ is assumed to be a point on the sphere such that the sum of squared geodesic distances between $\mathbf{R}^*$ and $\mathbf{R_1}, \ldots, \mathbf{R_N}$ is the minimum.

Summary of the Steps in Recognition: First the wb-scale scene $\mathcal{C}_{wb}$ of any test scene $\mathcal{C}$ is computed. Note that this does not require any explicit segmentation of the objects. From $\mathcal{C}_{wb}$, the PA system $\mathbf{PA}_{\text{intensity}}$ of the intensity structure is determined after using a fixed threshold. Then, from $\mathbf{F}$, the pose of the model assembly MA in $\mathcal{C}$ is determined from the relation $\mathbf{PA}_{\text{shape}} = (\mathbf{F})(\mathbf{PA}_{\text{intensity}})$. Once HSMOR has been completed, exact refinement gets done in the last step (delineation step) which is considered to be the fine level of recognition. In our experimental set up, we use the IGCASM strategy [22] to delineate 3-D structures, explained briefly in the next section.

C. Fine Recognition—Hybrid Segmentation

In our experimental set up, we use our previously described hybrid segmentation method, IGCASM [22], to delineate 3-D structures. In IGCASM, GC and ASM are combined synergistically to give better delineation accuracy than either method alone. In this study, we show both how accurate our proposed recognition platform is and how the recognition affects the final delineation. IGCASM effectively combines the rich statistical shape information embodied in 3-D ASM with the globally optimal delineation of Graph-Cut.[1] Once the MA is recognized in the coarse level, the delineation algorithm is used to finalize the whole segmentation process. Briefly, in addition to the traditional GC penalty terms (data and boundary penalty terms), a shape functional is integrated into the GC cost function in

[1]GC is a globally optimal segmentation method only for two-label segmentation. For the multilabel segmentation problem, it is a NP-hard problem. Although GC may not give a globally optimal segmentation result for multiobject segmentation, in IGCASM, we incorporate ASM with GC using the alpha-expansion method [24], which can find segmentation within a known factor of the global optimum.

IGCASM. Voxels inside or in the vicinity of the mean shape boundary are encouraged for the cut process and voxels outside and away from the model boundary are discouraged. This process is formulated with a shape functional similar to the data term in GC cost formulation, which is minimized through a conventional alpha-expansion method [24]. For parameter training and other technical details of the IGCASM method, see [22].

V. Evaluation and Results

A. Data

The performance of the proposed methodology has been evaluated on two datasets: 20 abdominal organs in low resolution CT images, and 11 foot MR images. The voxel size of the CT images is 1.17 mm $\times$ 1.17 mm $\times$ 1.17 mm (interpolated from 5 mm slices). Since our goal in this effort was to create models of normal anatomy, the participating radiologists reviewed and selected the images that were as close to normality as possible. Foot MRI data were acquired on a clinical 1.5T GE MRI machine, by using a coil specially designed for the study [34]. During each acquisition, the foot of the subject was locked in a nonmagnetic device. This allows the control of orientation and the motion of the foot. The imaging protocol used a 3-D steady-state gradient echo sequence with a $\mathrm{TR/TE/Flip\ angle} = 25\ \mathrm{ms}/10\ \mathrm{ms}/25^{\circ}$. The voxels are of size $0.55 \times 0.55 \times 0.55\ \mathrm{mm}^3$ (interpolated from slices 1.5 mm apart). The slice orientation was sagittal.

B. Ground Truth and Evaluation Criteria

We produced the ground truth data set for the CT and MRI volumes as described in Section III-A. For each subject, we generated a manually edited volume which labeled each voxel as being a particular object O_i, $1 \leq i \leq M$ (i.e., liver, spleen, talus, tibia, etc.) or background. Manual delineations were done slice by slice by experts using the Live-Wire algorithm [13]. Those 33 binary volumes (i.e., 20 CT and 11 MR images) with corresponding labels constitute our gold-standard data for our experiments and evaluations.

For each recognition experiment, we examine its accuracy and correctness by two validation methods: pose accuracy and delineation accuracy. We assess the proposed recognition algorithm's abilities for accurately locating the anatomical structures by leave-one-out-cross-validation (LOOCV) test. In order to assess the best recognition performance based on different combinations of structures in MA, we use all possible different combination of structures in the recognition experiments. Abbreviations and descriptions of those scenarios are listed in Tables I and II. Our aim was to better understand the advantage of using a large number of objects over single object recognition. Hence, we tried different scenarios where size and spatial position of the objects play an important role in recognition. Apart from recognition results, as a comparison and to be complete, we present also the delineation results of some particular scenarios.

1) A Down-Sampling Approach and its Sensitivity Analysis in WB-Scale Computation: While whole body CT images can take about 6 min in original resolution, abdominal images in original resolution can take a couple of minutes depending on the number of slices in the scene for wb-scale computation. If the image is down sampled by a factor of 4, the scale computation can be completed in 30 s. We observed the correlation between shape structure systems obtained using the original and down-sampled images to be $R = 0.9993$. Similarly, intensity structure systems obtained using the original and down-sampled grey level images yield a correlation of $R = 0.9899$. These results validate the use of down sampling to speed up wb-scale computation and still the construction of reliable relationship functions.

2) Evaluation of Scale Estimation: In the training step, the delineated objects are aligned in the seven-dimensional affine space as described previously. Owing to this alignment, the size differences within the subjects are uniformly handled. This leads to the *range of the scale component* in LOOCV tests to be a tight interval $(0.97 - 1.07)$. The scale range is obtained as follows: truly delineated shapes are enclosed by their minimum enclosing boxes. The scale range value of 1 then corresponds to the mean diagonal. In our experiments, we found errors in scale estimation to be 0.04 ± 0.017 and 0.025 ± 0.005 for abdominal and foot data, respectively.

3) Evaluation of Translation and Orientation Estimations: Fig. 5(a) and Fig. 11 (blue plot) show a summary of the recognition accuracies for different scenarios in terms of mean translation errors (MTEs) and standard deviation (SD) of MTEs over all subjects for the abdominal CT and foot MRI data, respectively. Scenarios are shown along the horizontal axis in all plots. The MTEs of foot MRI data are negligibly small, but the SD are not, as seen by red arrows. As readily noticed, the minimum MTEs and SD values are obtained when

TABLE I
Abbreviations of the Scenarios Used for Recognition and Their Corresponding Descriptions for Abdominal CT Dataset

Scenarios	Description
1-(LV)	Liver
2-(S)	Spleen
3-(LK)	Left Kidney
4-(RK)	Right Kidney
5-(LV+S)	Liver and Spleen
6-(LV+LK)	Liver and Left Kidney
7-(LV+RK)	Liver and Right Kidney
8-(LV+S+LK)	Liver, Spleen, and Left Kidney
9-(LV+S+RK)	Liver, Spleen, and Right Kidney
10-(S+LK)	Spleen and Left Kidney
11-(S+RK)	Spleen and Right Kidney
12-(S+LK+RK)	Spleen, Left Kidney, and Right Kidney
13-(LK+RK)	Left Kidney and Right Kidney
14-(LV+LK+RK)	Liver, Left Kidney and Right Kidney
15-(All)	Liver, Spleen, Left Kidney and Right Kidney

TABLE II
Abbreviations of the Scenarios and Their Corresponding Descriptions for Foot MRI Dataset. We Use the Following Symbols to Denote Foot Bones: Calcaneus: ca, Cuboid: cu, Navicular: na, Talus: ta, Tibia: ti

1:ca	2:cu	3:na	4:ta
5:ti	6:ca+cu	7:ca+na	8:ca+ti
9:ti+na	10:ti+ta	11:cu+na	12:ca+ta
13:na+ta	14:cu+ta	15:cu+ti	16:ca+na+ta
17: ca+ti+na	18:cu+na+ta	19:cu+ti+na	20:ca+cu+na
21:ca+cu+ta	22:ca+cu+ti	23:na+ta+ti	24:cu+ta+ti
25:ca+ta+ti	26:ca+cu+na+ta	27:ca+cu+ti+na	28:ca+cu+ti+ta
29:ca+na+ta+ti	30:cu+ti+na+ta	31:(all objects)	-

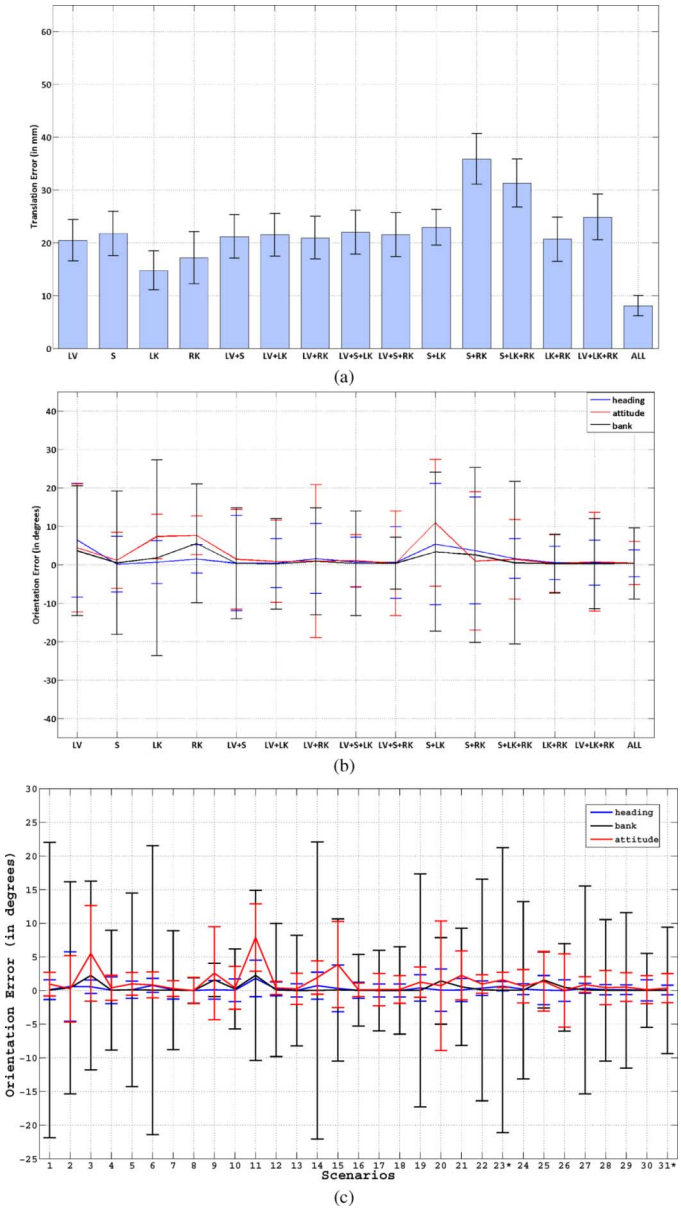

Fig. 5. Recognition accuracy in terms of MTEs (mm) and MOEs (in degrees) for abdominal CT data (a), (b) and foot data (c) with different number and combination of organs included in the model assembly MA. X: scenario (see Table I), Y: MTEs in mm or MOEs in degrees.

a large number of objects is included in the recognition process. Similarly, Fig. 5(b) and (c) shows recognition accuracy in terms of mean orientation errors (MOEs) (in degrees) and SD of MOEs over all subjects for the abdominal CT and foot MRI data, respectively. Note again that the minimum MOEs and SD values, computed separately in the directions of heading (x), attitude (y), and bank (z), are obtained if multiple objects are included in the recognition process. Interestingly, MOE in the direction of bank is higher compared to other directions. A possible reason for this is that the spatial resolution in the z direction is lower than in other directions. MOEs are about 10° if all objects are considered in recognition (scenario "all"). We point out that the best orientation accuracy is obtained when scenario "LK+RK" and "all" are used; that is, the combinations of left and right kidneys and all organs provide better orientation estimates. Furthermore, relatively lower recognition accuracy is obtained when spleen is included in the MA either with left kidney or right kidney alone. The reason behind this result has its basis in the variation of the organs' size, shape, and position considered in the MA. For example, spleen can vary in size, shape, and position based on the size and shape of the surrounding viscera, the position of which is dependent on how much the stomach is filled and the amount of blood in the spleen itself. Although most of these anatomic variants can be thought of as having no clinical significance, they need, however, to be recognized by the radiologists, as awareness of these variants is important to interpret the findings correctly and avoid mistaking them for a clinically significant abnormality.

Fig. 6 demonstrates the effectiveness of the proposed recognition method by displaying the original segmented abdominal organs (ground truths) in red, and the corresponding MAs in yellow in a series of scans.

4) Evaluation of Fine Level Recognition (Delineation): Following [28], we use the following accuracy measures for the quantitative evaluation of object delineation results. In order to characterize the delineation accuracy, the following two independent measures are defined: true positive volume fraction (TPVF), and false positive volume fraction (FPVF). (TPVF) and $(1 - \mathrm{FPVF})$ are defined as *delineation sensitivity and delineation specificity* of the segmentation method, respectively. In addition, we report dice similarity coefficients (DSC) for the delineation accuracies [45]. High values of those quantities indicate a good delineation accuracy. Table III lists the mean and SD values of delineation sensitivity, specificity, and DSC, over all objects and over the scene population, achieved in the two experiments by using the IGCASM^r method, where r indicates that HSMOR is applied to locate MA. As seen from Table III, IGCASM^r produces accurate delineations. All experiments have been performed on a Pentium 3.2 GHz PC with 2 GB RAM. While wb-scale filtering of a scene with dimension $512 \times 512 \times 150$ takes around 30 s, the average total time for the complete delineation of all objects takes about 39 s.

Table IV shows the mean and standard deviation values of specificity and sensitivity over all objects and over all abdominal CT and foot MRI data achieved in the two experiments by using IGCASM and IGCASM^r algorithms, where the difference between the two methods is solely due to the applied proposed initialization. In the IGCASM^r algorithm, the pose of the MA is estimated by using the proposed recognition method (scenario 15 and 31, respectively). In the IGCASM algorithm, MA is incorporated into the GC framework without using the proposed recognition method. IGCASM^r produces considerably more accurate delineations than the IGCASM method. It is clear that recognition is an important aspect of segmentation such that inappropriate initialization of the MA gives much lower segmentation accuracy. In addition, the best, average, and the worst segmentation results (based on the DSC values) for particular slices belonging to foot and abdominal images are illustrated in the first, second, and third columns of Fig. 7, respectively.

C. Comparison to Other Recognition Methods

In this experiment, an objective comparison between two well established organ localization methods (multiclass re-

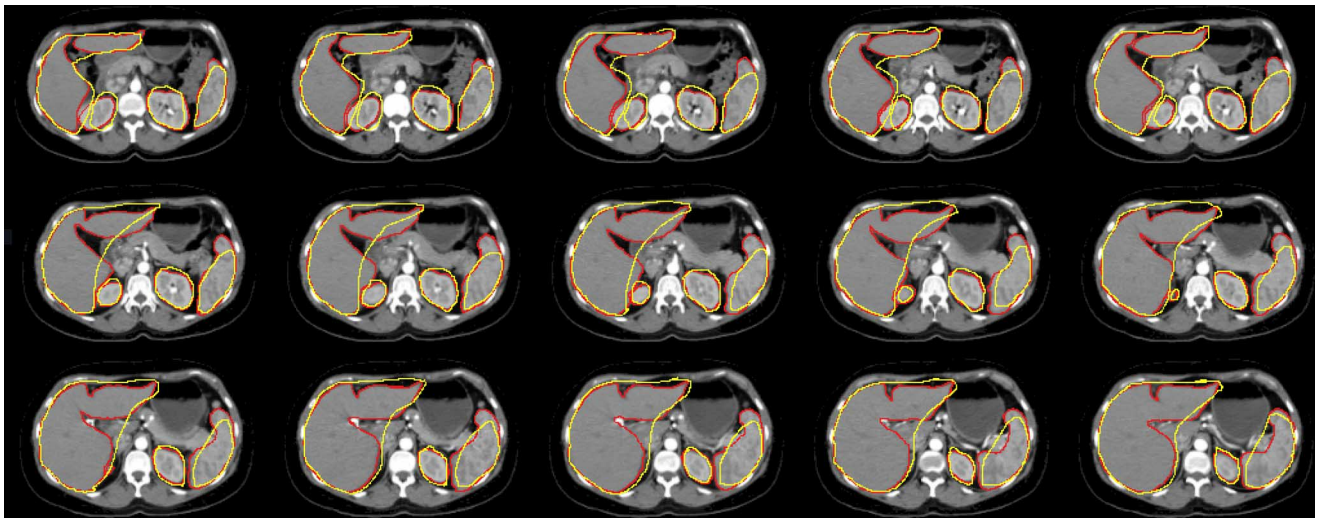

Fig. 6. Recognized MAs are shown in yellow while the ground truth segmentation of organs is shown in red.

TABLE III
MEAN AND SD OF (TPVF), $1 - \mathrm{FPVF}$, AND DSC FOR IGCASM^{r}

Abd. Organs	**TPVF**(%)	**1 − FPVF**(%)	**DSC**(%)
(Scn 1: Liver)	92.16±1.03	99.75±0.05	95.80±0.59
(Scn 2: Spleen)	93.47±1.28	99.77±0.07	96.51±0.80
(Scn 3: L.Kidney)	93.39±0.96	99.81±0.05	96.49±0.75
(Scn 4: R.Kidney)	93.55±0.92	99.80±0.03	96.57±0.71
(Scn 15: All organs)	93.01±1.05	99.78±0.05	96.27±0.82
Foot Bones	**TPVF**(%)	**1 − FPVF**(%)	**DSC**(%)
(Scn 1: Ca)	94.63±0.91	99.67±0.12	97.08±0.76
(Scn 2: Cu)	93.68±1.11	99.75±0.08	96.61±0.89
(Scn 3: Na)	93.17±1.29	99.74±0.07	96.34±1.02
(Scn 4: Ta)	94.89±0.97	99.73±0.09	97.24±0.78
(Scn 5: Ti)	92.36±1.27	99.72±0.06	95.89±0.95
(Scn 31: All bones)	93.75±1.11	99.72±0.08	96.63±0.89

TABLE IV
MEAN AND STANDARD DEVIATION OF (TPVF) AND $1 - \mathrm{FPVF}$ FOR IGCASM AND IGCASM^{r}

Data	**Method**	**TPVF**(%)	**1 − FPVF**(%)	**DSC**(%)
Abd. Organs	IGCASM	85.95 ± 8.64	99.87 ± 0.07	92.38±5.04
	IGCASM^{r}	93.01 ± 1.05	99.78 ± 0.05	96.27±0.82
Foot Bones	IGCASM	82.55 ± 8.76	99.60 ± 0.10	90.24±5.32
	IGCASM^{r}	93.75 ± 1.11	99.72 ± 0.08	96.63±0.89

gression forests [10] and atlas-based registration) and our HSMOR method is carried out over 20 abdominal CT scans. In the regression forests method, a direct nonlinear mapping is constructed from image space to organ location and size with training focusing on maximizing the confidence of output predictions [10]. We follow the steps described in [10] and use mean intensities over displaced, asymmetric cuboidal regions as visual features to capture spatial context. Similar to the study of [10], we use a fixed forest size of 12, and the maximum tree depth is found to be 7. In atlas-based registration methods [29], on the other hand, a reference template is constructed and localization is provided by registering any given test to the template. This process is summarized as follows: One of the scans in the training set is chosen as target scan randomly and all the scans in the training set are linearly aligned to that target

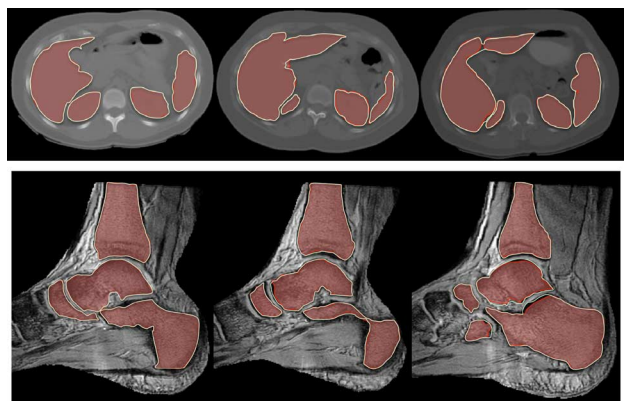

Fig. 7. Based on the DSC evaluation, the best (first column), average (second column), and worst (third column) segmentation results for particular slices are shown in white compared to the ground truth in red.

scan using a seven-parameter affine transformation. Second, we compute an intensity average template with a common position using the Define Common and Soft Mean modules of AIR software [29]. Third, we take intensity average template as target and repeat the above steps. This step is followed by a six-parameter rigid registration of all scans to the first target scan, resulting in the same spatial coordinate and scale of all scans in the training set. Fifth, we use a locally affine globally smooth registration method [27] to register all scans in the training set to the linear average template. Finally, we produce nonlinear template by computing an intensity average template from all linearly aligned scans including the target scan.

Our proposed HSMOR method achieved smaller MTEs compared to the regression forests and atlas based registration methods. Indeed, the SD of translation errors in the two methods is much higher compared to HSMOR. In overall organ localization, we achieve a MTE of less than 10 mm, and SD of MTEs below 2 mm. On the other hand, regression forests and atlas based methods have MTEs greater than 20 and 30 mm, respectively. The details of the MTEs and SD of MTEs are given in Fig. 8 for certain scenarios.

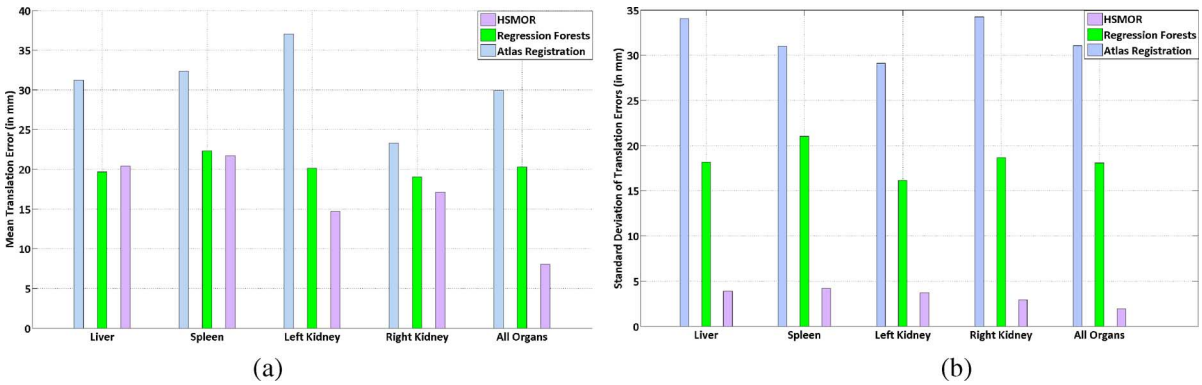

Fig. 8. MTEs and SD of MTEs are shown in (a) and (b), respectively.

VI. Intensity Nonstandardness Affects Recognition

In recognizing anatomical structures, we do not restrict ourselves only to the CT modality, but we also use MRI to show the robustness and generalizability of HSMOR. However, processing of MR images poses many challenges including the presence of noise, interpolation artefacts, intensity nonuniformities, and intensity nonstandardness. Among them, little attention has been paid to the effects of image intensity standardization/nonstandardness on image processing tasks. Since MR image intensities do not possess a tissue specific numeric meaning, even in images acquired for the same subject, on the same scanner, for the same body region, and obtained by using the same pulse sequence, it is important to transform the image scale into a standard intensity scale so that, for the same body region, intensities are similar. This process is called intensity standardization, a preprocessing technique mapping nonlinearly the image intensity grey scale of a given image into a standard intensity grey scale. In this section, we examine the role of intensity standardization in anatomy recognition tasks. In order to fully determine the effects of intensity nonstandardness on anatomy recognition, a controlled experimental framework is needed such that standardized and nonstandard images are both used in the recognition experiments for comparison purposes. To do so, first we need to obtain "clean" images, which do not include any inhomogeneities, intensity variations, or a high level of noise. Fig. 9 (A)–(H) illustrates the required experimental framework, following the study in [23]. "Clean" images are obtained through a series of operations: inhomogeneity correction (B), noise suppression (C), and standardization (D). All artefacts are removed from the images as best as possible so that only the effect of intensity standardization can be observed and studied. Once "clean" images are obtained, we add known levels of intensity nonstandardness to the "clean" images (E). The resulting images with different levels of nonstandardness are then used for anatomy recognition (F)–(H). This controlled framework allows us to determine to what extent intensity nonstandardness affects the recognition of anatomical structures. For nonuniformity correction (κ) and standardization (ψ), we use the method based on the concept of local morphometric scale called g-*scale* [33]. For noise suppression, a b-scale based diffusive filtering method was used such that the method preserves boundary sharpness and fine structures. For intensity standardization, we follow the steps reported in [23].

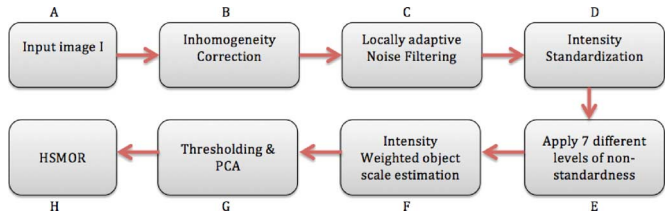

Fig. 9. Schematic illustration of the experimental framework for exploring the effects of intensity standardization on anatomy recognition.

A. Applying Nonstandardness

Let super-scripts, c, s, $\overline{s}$ and η denote, respectively, the scenes resulting from applying correction, standardization, introduction of nonstandardness, and scale-based diffusive filtering to a given scene. Thus the clean scene version of any scene $\mathcal{C} = (C, f)$ will be denoted $\mathcal{C}^{c\eta s} = (C, f^{c\eta s})$. To artificially introduce nonstandardness into a *clean scene* $\mathcal{C}^{c\eta s} = (C, f^{c\eta s})$, we use the idea of the inverse of the standardization mapping as described in [42]. Following [23], the intensities in a nonstandard scene $\mathcal{C}^{c\eta s\overline{s}} = \left(C, f^{c\eta s\overline{s}}\right)$ can be obtained by

$$f^{c\eta s\overline{s}}(\nu) = \begin{cases} \left\lceil \frac{f^{c\eta s}(\nu)}{m_1} \right\rceil, & \text{if } f^{c\eta s}(\nu) \le \mu_s \\ \left\lceil \frac{(f^{c\eta s}(\nu) - \mu_s)}{m_2} + \mu \right\rceil, & \text{if } f^{c\eta s}(\nu) > \mu_s \end{cases} \quad (5)$$

where $\lceil . \rceil$ converts any number $\mathrm{y} \in \mathbb{R}$ to the closest integer Y, μ_s denotes the median intensity on the standard scale, s_1, s_2 represent minimum and maximum intensity levels, and m_1 and m_2 denote the varying slopes (see [23] on how to estimate those parameters) shown in Fig. 10. We combine eight different ranges of the slopes m_1 and m_2, to introduce small, medium, and large scale nonstandardness. This means that, for each *clean scene*, we obtain eight scenes, one of which is the default *clean scene* itself, two scenes consisting of small scale nonstandardness, two scenes consisting of medium scale nonstandardness, and three scenes consisting of large scale nonstandardness. The ranges of the applied nonstandardness are summarized in Table V.

B. Compute WB-Scale Scenes

After different artificial nonstandardness with levels from $\overline{\psi}_1$ to $\overline{\psi}_7$ are added into the grey-level clean scenes, we use the IWOSE algorithm to compute wb-scale scenes.

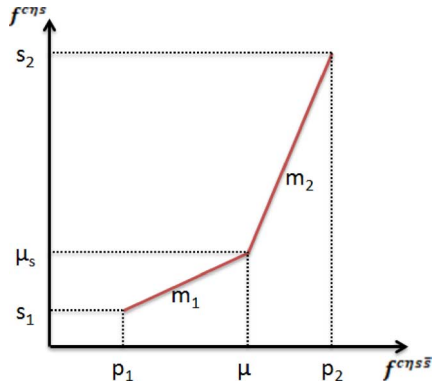

Fig. 10. Standardization transformation function for inverse mapping with the various parameters shown.

TABLE V
DESCRIPTION OF THE DIFFERENT RANGES OF THE SLOPES m_1, m_2 FOR INTRODUCING ARTIFICIAL NON-STANDARDNESS

function	Range	Description
$\overline{\psi}_1$	$\{0.9 \leq m_1, m_2 \leq 1.5\}$	Small Scale
$\overline{\psi}_2$	$\{0.6 \leq m_1, m_2 \leq 0.9\}$	
$\overline{\psi}_3$	$\{1.5 \leq m_1, m_2 \leq 2.0\}$	Medium Scale
$\overline{\psi}_4$	$\{2.0 \leq m_1, m_2 \leq 2.4\}$	
$\overline{\psi}_5$	$\{2.4 \leq m_1, m_2 \leq 2.7\}$	Large Scale
$\overline{\psi}_6$	$\{2.7 \leq m_1, m_2 \leq 3.0\}$	
$\overline{\psi}_7$	$\{3.0 \leq m_1, m_2 \leq 3.3\}$	

C. Determining Intensity and Shape Structure Systems

We apply a fixed threshold interval $thrs$ to the scenes derived through $\overline{\psi}_0$ to $\overline{\psi}_7$. We determine the intensity structure systems from the twb-scale scenes. Then, the relationships between the intensity and shape structure systems are modelled. Each intensity structure system constitutes a different relationship: $F, F_1, \ldots, F_7$, where F stands for the relationship function between $\mathbf{PA}^{\mathbf{clean}}_{\mathrm{intensity}}$ and $\mathbf{PA}_{\mathrm{shape}}$, and $F_1, \ldots, F_7$ stand for the relationship functions between $\mathbf{PA}^{\mathbf{1}}_{\mathrm{intensity}}$ and $\mathbf{PA}_{\mathrm{shape}}, \ldots, \mathbf{PA}^{\mathbf{7}}_{\mathrm{intensity}}$ and $\mathbf{PA}_{\mathrm{shape}}$, respectively.

D. Evaluation of Single and Multiobject Recognition Strategies

We use the relationship functions $F, F_1, \ldots, F_7$ for quick positioning of the MA in any given test image. Since estimation of the scale parameter is done in the training step from the delineated objects, and only a range of scale information is provided for the scale parameter selection, there is no scale difference between standardized and nonstandard scenes. Thus, the influence of nonstandardness on recognition includes only orientation and translation errors. We use LOOCV to measure recognition performance considering the seven different levels of nonstandardness together with one level of standardness (i.e., total of eight levels) and 31 different recognition scenarios in relation to the different combinations of the five different structures in the foot data.

The results of the comparison experiments of recognition for the scenario 31 (i.e., when all objects are used) are reported in Table VI for seven sets of nonstandard scenes derived from $\overline{\psi}_1, \ldots, \overline{\psi}_7$ with respect to the recognition performance of clean scenes derived from $\overline{\psi}_0$. The table summarizes MTEs (in mm), MOEs in heading (x), attitude (y), and bank (z) directions (in degrees), and their corresponding SD values. The ability to recognize objects is lower if the scenes include high levels of nonstandardness. A reason for the better recognition performance of clean scenes compared to the nonstandard scenes is that the fixed thresholding interval $thrs$ gives narrower limits for the pose parameters that describe the relationship of the model assembly MA to the intensity appearance. Fig. 11 shows recognition accuracy in terms of MTEs for different numbers and combination of structures included in the model assembly MA. For simplicity, we compare the recognition accuracy of scenes with only a high level of nonstandardness ($\overline{\psi}_7$) with respect to the recognition with clean scenes. As seen from the figure, almost for all cases, the recognition accuracy of standardized scenes wins over the recognition accuracy of nonstandard scenes. When nonstandardness is introduced into the clean scenes, relationship functions are affected nonlinearly because the introduction of nonstandardness is itself a nonlinear process. As the relationship functions are distorted nonlinearly, the solution space for the pose estimate of MA becomes large.

VII. CONCLUDING REMARKS

We observed that the effectiveness of object recognition depends on the number and distribution of objects considered in the model assembly. Recognition accuracy improves with increasing number of objects. The evaluated results indicate the following. 1) High recognition accuracy can be achieved by including a large number of objects which are spread out in the body region. 2) Incorporating local object scale information improves the recognition in a way that there is no need to do search for scaling, orientation, and translation parameters. That is the pose of objects can be estimated in one shot without search or optimization. 3) The appearance information incorporated via ball-scale has a strong effect on the computation of the PA system, and on the relationship function $\mathbf{F}$. 4) The incorporation of shape prior into the GC framework by embedding proper scale, orientation, and translation information is feasible. 5) Intensity variation among scenes in an ensemble degrades recognition performance, because it affects the relationship functions between shape and intensity structure systems. Specifically, the spread of the pose parameters increases considerably when scenes have intensity nonstandardness.

Further improvements on anatomy recognition may be perhaps gained if texture uniformity or Marginal Space Learning based features [11] are considered instead of the simple image intensity uniformity for estimating the ball scales. In this case, the specification of scale and all ensuing information can be made specific to the different image modalities (CT, MRI, US).

In this paper, we have not addressed the issue of handling abnormalities due to diseases or treatment. We believe that modelling should be (and perhaps can be) done only of normality, and through its knowledge, abnormality should be detected and delineated in given patient images. This is a topic of our current research.

TABLE VI
MEAN AND (SD) OF THE ORIENTATION AND TRANSLATION ERRORS FOR THE SCENARIO 31 OF FOOT MRI DATA ARE LISTED. THE TYPE OF NON-STANDARDNESS IS INDICATED BY $\overline{\psi}_0, \ldots, \overline{\psi}_7$, WHERE $\overline{\psi}_0$ DENOTES THAT THERE IS NO NON-STANDARDNESS APPLIED TO THE SCENE, NAMELY THE SCENE IS CLEAN

	$\overline{\psi}_0$	$\overline{\psi}_1$	$\overline{\psi}_2$	$\overline{\psi}_3$	$\overline{\psi}_4$	$\overline{\psi}_5$	$\overline{\psi}_6$	$\overline{\psi}_7$
MOE in x (deg)	0.0292	0.0925	0.0898	0.0615	0.0846	0.0869	0.1945	0.2094
SD in x (deg)	0.7088	0.7399	0.7321	0.7638	0.7684	0.7693	0.7562	0.7303
MOE in y (deg)	0.3576	0.3426	0.3840	0.3858	0.3899	0.3846	0.3528	0.3839
SD in y (deg)	2.0739	2.1311	2.2962	2.3018	2.3970	2.5217	2.3428	2.3060
MOE in z (deg)	0.0209	0.0264	0.0236	0.0341	0.0550	0.0266	0.0984	0.1157
SD in z (deg)	9.4049	9.7597	9.7541	9.7741	9.7796	9.7823	9.7324	9.7051
MTE in mm	2.1004	2.4500	2.5701	2.9700	3.4050	3.2000	3.6606	3.5709
SD in Trans. (in mm)	4.2247	4.2549	4.3690	4.2783	4.4600	4.5811	8.0536	9.7181

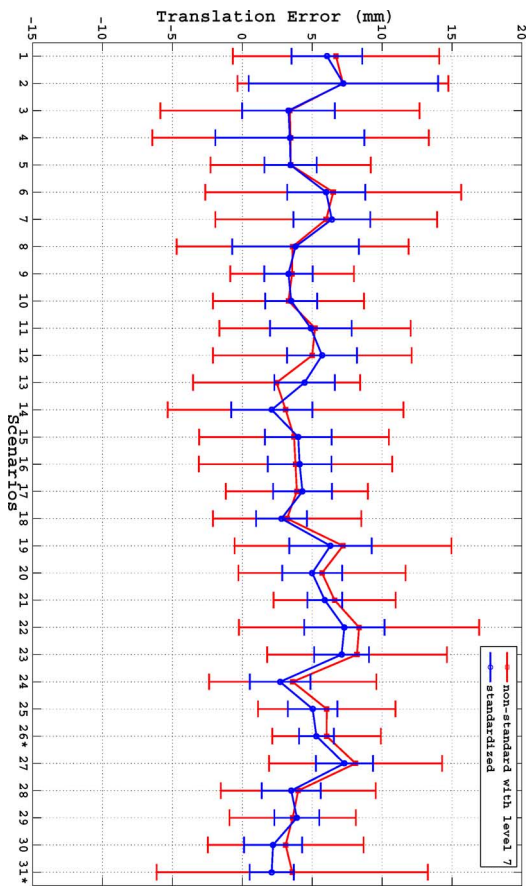

Fig. 11. Recognition accuracy in terms of MTEs (in mm) for foot MR images with different numbers and combination of structures included in the model assembly (see Table II for scenarios).

ACKNOWLEDGMENT

The authors would like to thank Dr. D. Torigian of the Department of Radiology, University of Pennsylvania, Dr. B. Hirsch of the Department of Neurobiology and Anatomy, Drexel University, for providing the data and helping in constructing ground-truth.

REFERENCES

[1] S. M. Pizer *et al.*, "Deformable m-reps for 3-D medical image segmentation," *Int. J. Comput. Vis.*, vol. 55, no. 2/3, pp. 851–865, 2003.
[2] J. Weese, M. Kaus, C. Lorenz, S. Lobregt, R. Truyen, and V. Pekar, "Shape constrained deformable models for 3-D medical image segmentation," in *Proc. Inf. Process. Med. Imag. (IPMI)*, 2001, vol. 2082, pp. 380–387.
[3] L. Soler *et al.*, "Fully automatic anatomical, pathological, and functional segmentation from CT scans for hepatic surgery," *Comput. Aided Surg.*, vol. 6, no. 3, pp. 131–142, 2001.
[4] M. Brejl and M. Sonka, "Object localization and border detection criteria design in edge-based image segmentation: Automated learning from examples," *IEEE Trans. Med. Imag.*, vol. 19, no. 10, pp. 973–985, Oct. 2000.
[5] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—Their training and application," *Comput. Vis. Image Understand.*, vol. 61, pp. 38–59, 1995.
[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 68–685, 2001.
[7] J. Fripp, S. Crozier, S. Warfield, and S. Ourselin, "Automatic initialisation of 3-D deformable models for cartilage segmentation," in *Proc. Digital Image Comput.: Tech. Appl.*, 2005, pp. 513–518.
[8] R. Bajcsy, R. Lieberson, and M. Reivich, "A computerized system for the elastic matching of deformed radio-graphic images to idealized atlas images," *J. Comput. Assist. Tomogr.*, vol. 7, no. 4, pp. 618–625, 1983.
[9] R. Bajcsy and A. Kovacic, "Multi-resolution elastic matching," *Comput. Graphics Image Process.*, vol. 46, pp. 1–21, 1989.
[10] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu, "Regression forests for efficient anatomy detection and localization in CT studies," in *MICCAI-MCV Workshop*, 2010.
[11] Y. Zheng *et al.*, "Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT volume using marginal space learning and steerable features," *IEEE Trans. Med. Imag.*, vol. 27, no. 11, pp. 1668–1681, Nov. 2008.
[12] D. G. Kendal, "A survey of statistical theory of shape," *Stat. Sci.*, vol. 4, pp. 87–120, 1989.
[13] A. X. Falcao *et al.*, "User-steered image segmentation paradigms: Live wire and live lane," *Graph. Models Image Process.*, vol. 60, no. 4, pp. 233–260, 1998.
[14] D. G. Altman, *Practical Statistics for Medical Research*. London, U.K.: Chapman Hall, 1991.
[15] K. V. Mardia and I. L. Dryden, "The statistical analysis of shape data," *Biometrika*, vol. 76, no. 2, pp. 271–281, 1989.
[16] R. Davies, C. Twining, and C. Taylor, *Statistical Models of Shape: Optimisation and Evaluation*, 1st ed. New York: Springer, 2008.
[17] C. G. Small, *The Statistical Theory of Shape*. New York: Springer, 1996.
[18] U. Bagci, J. K. Udupa, and X. Chen, "Ball-scale based multi-object recognition in a hierarchical framework," in *Proc. SPIE Med. Imag.*, 2010, vol. 7623, pp. 762345-1–762345-12.
[19] P. K. Saha, J. K. Udupa, and D. Odhner, "Scale-based fuzzy connected image segmentation: Theory, algorithms, and validation," *Comput. Vis. Image Understand.*, vol. 77, pp. 145–174, 2000.
[20] P. K. Saha and J. K. Udupa, "Scale-Based diffusive image filtering preserving boundary sharpness and fine structures," *IEEE Trans. Med. Imag.*, vol. 20, no. 11, pp. 1140–1155, Nov. 2001.
[21] L. Nyul, J. K. Udupa, and P. K. Saha, "Incorporating a measure of local scale in voxel-based 3-D image registration," *IEEE Trans. Med. Imag.*, vol. 22, no. 2, pp. 228–237, Feb. 2003.
[22] X. Chen, J. K. Udupa, U. Bagci, A. Alavi, and D. A. Torigian, "3-D automatic anatomy recognition based on iterative graph-cut-ASM," in *Proc. SPIE Med. Imag.*, 2010, vol. 7625, pp. 76251T-1–76251T-8.
[23] U. Bagci, J. K. Udupa, and L. Bai, "The role of intensity standardization in medical image registration," *Pattern Recognit. Lett.*, vol. 31, no. 4, pp. 315–323, 2010.
[24] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph-cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1222–1239, Dec. 2001.
[25] P. K. Saha, "Tensor scale: A local morphometric parameter with applications to computer vision and image processing," *Comput. Vis. Image Understand.*, vol. 99, no. 3, pp. 384–413, 2005.

[26] K. R. Castleman, *Digital Image Processing*. Englewood Cliffs, NJ: Prentice Hall, 1996.
[27] U. Bagci and L. Bai, "Automatic best reference slice selection for smooth volume reconstruction of a mouse brain from histological slices," *IEEE Trans. Med. Imag.*, vol. 29, no. 9, pp. 1688–1696, Sep. 2010.
[28] J. K. Udupa *et al.*, "A framework for evaluating image segmentation algorithms," *Computerized Med. Imag. Graphics*, vol. 30, no. 2, pp. 75–87, 2006.
[29] Y. Tang *et al.*, "The construction of a Chinese MRI brain atlas: A morphometric comparison study between Chinese and Caucasian cohorts," *NeuroImage*, vol. 51, no. 1, pp. 33–41, 2010.
[30] M. P. Kumar, P. H. S. Torr, and A. Zisserman, "OBJCUT: Efficient segmentation using top-down and bottom-up cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 530–545, Mar. 2010.
[31] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graphics (SIGGRAPH)*, vol. 23, pp. 309–314, 2004.
[32] P. K. Saha and J. K. Udupa, "Optimum image thresholding via class uncertainity and region homogeneity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 7, pp. 689–706, Jul. 2001.
[33] A. Madabhushi, J. K. Udupa, and A. Souza, "Generalized scale: Theory, algorithms, and application to image inhomogeneity correction," *Comput. Vis. Image Understand.*, vol. 101, no. 2, pp. 100–121, 2006.
[34] J. K. Udupa, B. E. Hirsch, H. J. Hillstrom, G. R. Bauer, and J. B. Kneeland, "Analysis of in vivo 3-D internal kinematics of the joints of the foot," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 11, pp. 1387–1396, Nov. 1998.
[35] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient N-D image segmentation," *Int. J. Comput. Vis.*, vol. 70, no. 2, pp. 109–131, 2006.
[36] Y. Boykov and V. Kolmogorov, "Computing geodesics and minimal surfaces via graph cuts," in *Proc. ICCV*, 2003, pp. 26–33.
[37] I. L. Dryden and K. V. Mardia, *Statistical Shape Analysis*. New York: Wiley, 1998.
[38] P. K. Saha and J. K. Udupa, "Optimum image thresholding via class uncertainity and region homogeneity," *IEEE Trans. Pattern Ana. Mach. Intell.*, vol. 23, no. 7, pp. 689–706, Jul. 2001.
[39] A. Rangarajan, H. Chui, and F. L. Bookstein, "The softassign procrustes matching algorithm," in *Proc. IPMI*, 1997, vol. 1230, pp. 29–42.
[40] A. Kelemen, G. Szekely, and G. Gerig, "Elastic model-based segmentation of 3-D neuroradiological data sets," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 828–839, Oct. 1999.
[41] B. Tsagaan, A. Shimizu, H. Kobatake, and K. Miyakawa, "An automated segmentation method of kidney using statistical information," in *Proc. MICCAI*, 2002, vol. 2488, pp. 556–563.
[42] A. Madabhushi and J. K. Udupa, "Interplay between intensity standardization and inhomogeneity correction in MR image processing," *IEEE Trans. Med. Imag.*, vol. 24, no. 5, pp. 561–576, May 2005.
[43] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3-D medical image segmentation: A review," *Med. Image Anal.*, vol. 13, no. 4, pp. 543–563, 2009.
[44] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Shape matching and anisotropy," in *ACM Trans. Graphics (Proc. SIGGRAPH)*, Aug. 2004.
[45] X. Zhuang, K. Leung, K. Rhode, R. Razavi, D. Hawkes, and S. Ourselin, "Whole heart segmentation of cardiac MRI using multiple path propagation strategy," in *Proc. MICCAI*, 2010, vol. 13(1), pp. 435–443.