Performance benchmarking of liver CT image segmentation and volume estimation

Wei Xiong [a], Jiayin Zhou [b], Qi Tian [a], Jimmy J. Liu [a], Yingyi Qi [c], Wee Kheng Leow [c]
Thazin Han [b], Shih-chang Wang [b]
[a] Institute for Infocomm Research, A*STAR, Singapore, {wxiong, tian, jliu@i2r.a-star.edu.sg}
[b] School of Medicine, Nat'l Univ. of Singapore, Singapore {dnrzjy, dnrth, dnrhead@nus.edu.sg}
[c] School of Computing, Nat'l Univ. of Singapore, Singapore, {qiyingyi, leowwk@comp.nus.edu.sg}

ABSTRACT

In recent years more and more computer aided diagnosis (CAD) systems are being used routinely in hospitals. Image-based knowledge discovery plays important roles in many CAD applications, which have great potential to be integrated into the next-generation picture archiving and communication systems (PACS). Robust medical image segmentation tools are essentials for such discovery in many CAD applications. In this paper we present a platform with necessary tools for performance benchmarking for algorithms of liver segmentation and volume estimation used for liver transplantation planning. It includes an abdominal computer tomography (CT) image database (DB), annotation tools, a ground truth DB, and performance measure protocols. The proposed architecture is generic and can be used for other organs and imaging modalities. In the current study, approximately 70 sets of abdominal CT images with normal livers have been collected and a user-friendly annotation tool is developed to generate ground truth data for a variety of organs, including 2D contours of liver, two kidneys, spleen, aorta and spinal canal. Abdominal organ segmentation algorithms using 2D atlases and 3D probabilistic atlases can be evaluated on the platform. Preliminary benchmark results from the liver segmentation algorithms which make use of statistical knowledge extracted from the abdominal CT image DB are also reported. We target to increase the CT scans to about 300 sets in the near future and plan to make the DBs built available to medical imaging research community for performance benchmarking of liver segmentation algorithms.

Keywords: PACS, Abdominal CT imaging, Liver Segmentation, Performance Benchmark, Knowledge Extraction, Probabilistic Atlas, CAD

1. INTRODUCTION

In recent years more and more computer aided diagnosis (CAD) systems are being used routinely in many hospitals and specialist centers. Image-based knowledge discovery plays important roles in many CAD applications, which have great potential to be integrated into the next-generation picture archiving and communication systems (PCAS)[1,2]. In order to extract knowledge from large volumes of medical images stored in PACS systems, medical image databases (DBs) have to be constructed based on needs from specific applications and diseases. During the past years, research efforts have been devoted to build medical image DBs for CAD and content based medical image retrieval (CBMIR), such as the Lung Imaging Database Consortium (LIDC) for lung cancer diagnosis in USA[3] and imageCLEF for CBMIR in Europe.[4,5] Such DBs have been greatly supporting the research community.

In general, a typical image-based CAD procedure includes the isolation of region of interest (ROI), feature extraction, pattern analysis and interpretation, and decision-making. Hence, efficient and robust medical image segmentation tools are essential in many CAD applications. With established DBs for specific applications, medical images will be processed and segmented for low-level feature extraction and ground truth data will be prepared. High-level organ knowledge, such as 2D / 3D atlases as well as pathological features, can be computed based on the medical image data and the ground truth built. Once these collections of data and knowledge are available, they can then be used to benchmark various tools and systems for medical image segmentation, classification and retrieval.

Research efforts have been put into liver segmentation and volume estimation during the past decade.[6-9] In order to meet clinical needs, accurate and robust abdominal organ segmentation is still very challenging due to 1) complex anatomic layouts and very similar densities for different organs in abdominal region, and 2) large variations in shape and locations of the same organs among humans. Due of the limited amount of liver CT data sets used for each study and the

Medical Imaging 2008: PACS and Imaging Informatics, edited by Katherine P. Andriole, Khan M. Siddiqui,
Proc. of SPIE Vol. 6919, 69190N, (2008) · 1605-7422/08/$18 · doi: 10.1117/12.770858

fact that they are not available to other researchers, it is hard to compare performance among different liver segmentation algorithms. Hence it is well-expected to have open source of medical image DBs with associated ground truth and common platforms for the community to benchmark various segmentation algorithms, applied to various organs and imaging modalities. A workshop was organized in MICCAI 2007 for the competition of caudate and liver segmentations.[10] In this workshop, a small MRI brain image DB and a small CT liver image DB containing both normal and pathological cases, were provided for training (with reference segmentations), testing and validation. Currently, however, there is still no publicly available large DB of abdominal CT images for the research community.

The purpose of the present work is to propose knowledge-extraction architecture from medical images for 3D abdominal organ segmentation and volume estimation in the context of CAD towards the next generation PCAS systems. With such architecture, algorithms will be benchmarked for 3D liver segmentation and volume estimation. The proposed architecture is however generic and can be used for other organs such as kidney, pancreas, etc.

The organization of this paper is as follows: Section 2 gives an overview of this architecture platform with briefing and figure illustration. Necessary steps in constructing the architecture including data recruitment, data labeling and statistical knowledge extraction are introduced in Section 3. In Section 4 the performance benchmarking of liver segmentation algorithms using the constructed abdominal CT image DB is presented. Results and discussion are described in Section 5 and the paper is concluded lastly.

2. PLATFORM OVERVIEW

A generic platform for benchmarking liver image segmentation algorithms and volume estimation is shown in Fig. 1, It includes following components: an abdominal CT image DB, a ground truth DB, annotation tools, performance measure protocols, and finally, statistical models extracted from image and ground truth DBs, such as 2D/3D atlases for individual and multiple organs. There are two phases involved to construct the platform and the DBs. Phase 1 consists of establishing an abdominal CT image DB; preparing ground truth from medical images for individual organs; and computing statistical models such as probabilistic atlas for liver, kidneys and spleen, etc., based on images and the ground truth DBs. This procedure is associated with different registration and alignment algorithms and the computed statistical atlas will be used as the prior knowledge to facilitate further organ segmentation. Phase 2 involves testing various segmentation algorithms against the CT image DB; comparing obtained results with the ground truth data; and calculating performance indexes based on given performance protocols. Phase 1 is considered as a learning phase while Phase 2 is considered as a testing phase.

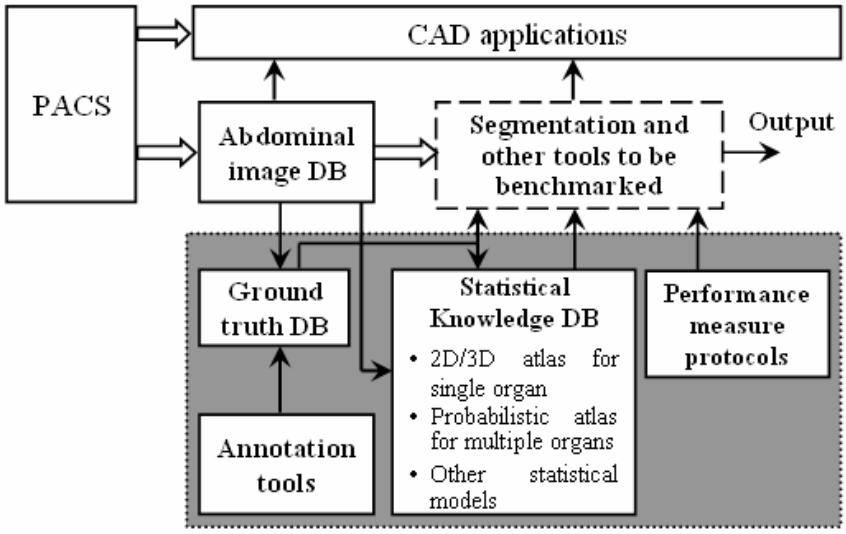

Figure 1. The proposed generic platform with image-based knowledge extraction tools and the application in 3D liver CT image segmentation and volume

3. DATABASE CONSTRUCTION AND KNOWLEDGE EXTRACTION

3.1 Image Recruitment and Ground Truth Labeling

We have been granted by a local ethics committee to access human abdominal CT images for our research purpose. In the first stage, 70 sets of scans imaged by a spiral scanner (Volume Zoom, Siemens, Germany) are collected from the Department of Diagnostic Imaging, National University Hospital, Singapore. The image recruitment criteria are as follows: 1). Axial images were acquired using a standard clinical abdominal protocol from upper abdomen to pelvic sections, 5 mm slice thickness without inter-slice gap; 2). Non-ionic contrast enhanced scan at the arterial phase; 3). No focal lesion is seen in liver, spleen and kidneys but fatty liver is allowed. Eventually about 300 abdominal CT scans including normal and pathological cases will be collected into the DB.

After data recruitment, 2D ground truth contours of liver, kidneys, spleen, aorta and spinal canal in each data set were manually traced on the basis of each slice by using an interactive pen-display (DTI520, Wacom, Japan) and a user-friendly annotation tool developed by this group, as shown in Fig. 2. The annotation process was carried out by one radiologist and one imaging scientist and the boundaries traced were under their mutual agreement. Fig. 3 shows a CT image with traced ground truths for organs and structures, and Fig. 4 shows a 3D view of a traced liver.

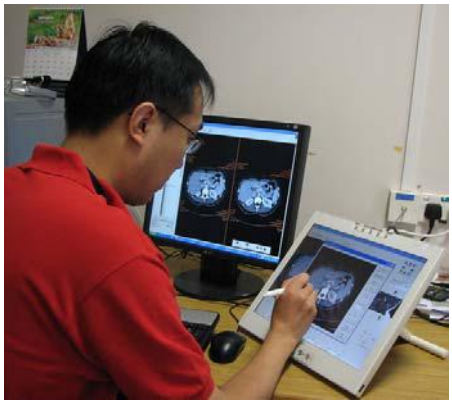

Figure 2. The image annotation tool with an interactive pen-display

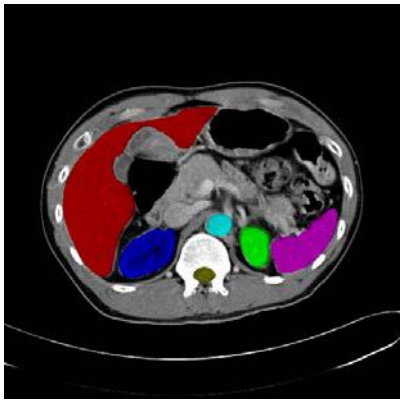

Figure 3. A CT image with traced and labeled organs and structures

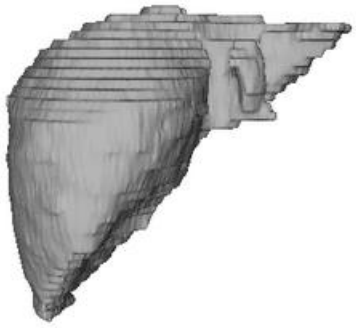

Figure 4. 3D visualization of a traced liver

3.2 Construction of 2D Contour Atlases of Abdominal Organs

2D contour atlases are constructed from the ground truth data, and they will be used as the initialization of two liver segmentation algorithms to be benchmarked. We construct a simple 2D atlas from only one slice of abdominal CT image by extract its contour. This is show in Fig. 5 where the annotated liver CT image (the left) is thresholded to obtain the region containing the liver (the middle) and traced contour (the right). We call this atlas as a single-2D atlas. We have also constructed a composite 2D contour atlas from multiple slices of images. Fig. 6 shows three gray-level CT images (the 3 left images) used to construct a single probabilistic atlas. The three images are firstly thresholded and then their logic OR is applied to obtained a region containing possible locations where a liver pixel can occur (the 4th from left). Its contour is then considered as a composite-2D atlas (the right).

Figure 5. Construction of a single 2D atlas using single slice of annotated image

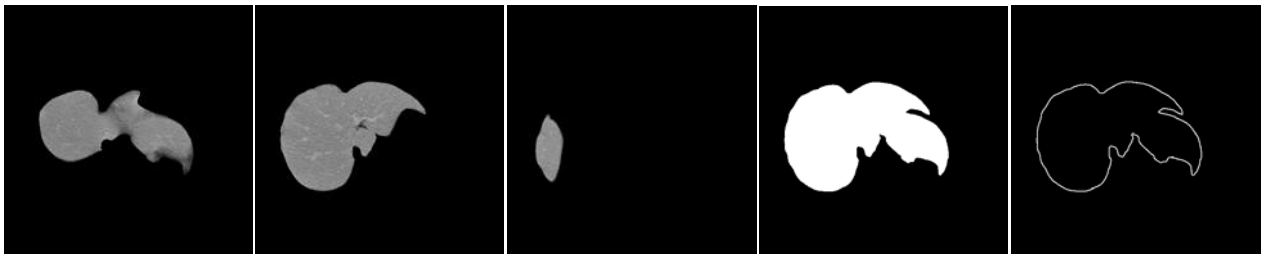

Figure 6. Construction of a composite 2D atlas from multiple annotated 2D images

3.3 Construction of a Probabilistic Atlas of Abdominal Organs

Probabilistic atlas refers to a set of maps showing "the complete spatial distribution of probabilities that a voxel belongs to one or more organs". [11] Each voxel is a vector whose cardinality depends on the number of organs. It is believed that probabilistic atlas can introduce more possible spatial information than single and deterministic atlas to the definition of complex and deformable organs. The basic idea to construct this atlas is to choose one common target set from some labeled abdominal data sets, register other sets to this target, and count, for each class of organs, how many voxels in these sets will be mapped to each voxel in the target. [12] Once a set of corresponding landmarks are chosen in both the reference and the target sets, the warping transform maps each voxel in the reference set to the target set and an objective function is computed to measure the goodness of the set of landmarks. An optimization approach will move the landmarks in the reference set to initialize an iterative procedure to map the reference set and compute the objective function again till the function is minimized. The warping transform used here is the thin-plate splines (TPS).[13] The objective function is based on the mutual information (MI) between the target set and its reference set; and the multivariate nonlinear optimization algorithm to optimize the registration is a downhill simplex method.[14]

3.3.1 Thin-plate splines and MI-based registration

The warping method using thin-plate splines interpolates surfaces over irregular spacing data by minimizing a physical bending energy functional of a thin metal plate using landmark point constraints. The two-dimensional interpolation applications of TPS were pioneered by Bookstein[13] and here we adapt an extension to three dimensions.[15]

Given n pairs of landmarks $\{x_i, y_i, z_i\}_{i=1}^n$ and $\{x'_i, y'_i, z'_i\}_{i=1}^n$ in the reference set and the target set respectively, the thin plate spline maps any point $\mathbf{x} = (x, y, z)$ in the reference space to $\mathbf{x'} = (x', y', z')$ in its target space:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \mathbf{A}\begin{bmatrix} 1 & x & y & z \end{bmatrix} + \sum_{i=1}^{n} w_i U(|\mathbf{x} - \mathbf{x}_i|) . \tag{1}$$

Here $U(|\mathbf{x} - \mathbf{x}_i|)$ is the three-dimensional Cartesian distance between $\mathbf{x}$ and $\mathbf{x_i}$. $\mathbf{A} = \begin{bmatrix} a_1 & a_x & a_y & a_z \end{bmatrix}^T$ and $\mathbf{W} = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}^T$ are TPS model parameters and are determined as follows. Let $U(r_{ij}) = U(|\mathbf{x}_i - \mathbf{x}_j|)$, $\mathbf{O}_{4\times4}$ and $\mathbf{O}_{3\times4}$ be a 4-by-4 and a 3-by-4 matrix of zeros, respectively. Denote

$$\mathbf{V} = \begin{bmatrix} x'_1 & x'_2 & \cdots & x'_n \\ y'_1 & y'_2 & \cdots & y'_n \\ z'_1 & z'_2 & \cdots & z'_n \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 1 & x_1 & y_1 & z_1 \\ 1 & x_2 & y_2 & z_2 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & y_n & z_n \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} 0 & U(r_{12}) & \cdots & U(r_{1n}) \\ U(r_{21}) & 0 & \cdots & U(r_{2n}) \\ \cdots & \cdots & \cdots & \cdots \\ U(r_{n1}) & U(r_{n2}) & \cdots & 0 \end{bmatrix}, \quad \text{and} \quad \mathbf{L} = \begin{bmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P} & \mathbf{O}_{4\times4} \end{bmatrix},$$

$\mathbf{Q} = \begin{bmatrix} \mathbf{V} & \mathbf{O}_{3\times4} \end{bmatrix}^T$. The j-th column vector of $\mathbf{Q}$ is a (n+4)-vector $\mathbf{Q}_j$, $j = 1, 2, 3$. The model parameters $\mathbf{A}$, $\mathbf{W}$, corresponding to $\mathbf{Q}_j$, are found by $\begin{bmatrix} \mathbf{W} & \mathbf{A} \end{bmatrix}^T = \mathbf{L}^{-1}\mathbf{Q}_j$, $j = 1, 2, 3$. Given two data sets, A and B, and their marginal probability density functions $p(A)$, $p(B)$ and their joint probability density function $p(A,B)$, the mutual information between them is defined as

$$MI(A,B) = \sum_{A}\sum_{B} p(A,B)\log_2(p(A,B)/(p(A)(p(B))) . \quad (2)$$

Note that the summations are over all possible feature values. The dissimilarity between them can be defined as $-MI(A,B)$. We use the Nelder-Mead downhill simplex method[14] for our multivariate nonlinear optimization. As we are handling 3D volumetric data, the dimension for the optimization problem is $3n$ for n landmarks.

3.3.2 Construction of a multiple-organ probabilistic atlas

We register a group of labeled reference sets to one common labeled target set. Four landmarks are labeled in each set of volume data: two in the liver, one in the left kidney and one in the right kidney. For each registration, the optimizer drives the four points in the reference set so that a local minimum of objective function is achieved. A TPS model is then found using these optimized landmarks and using this model the whole set of points are mapped to the targets set.

During this warping transform, at each voxel $\mathbf{x}' = (x', y', z')$ in the target space, every occurrence of organ class $c = 1,...,C$, is counted, yielding a vector C components. For class c, let $\Pr(c \mid \mathbf{x}')$ be its conditional probability and $\mathbf{G}_c$ its certain attribute, then the attribute for the final warped volumes at $\mathbf{x}'$ will be

$$\mathbf{G} = \sum_{c=1}^{C} \Pr(c \mid \mathbf{x}')\mathbf{G}_c . \quad (3)$$

Once all voxels in the target space are determined, the probabilistic atlas is constructed.Suppose, at $\mathbf{x}'$, the number of counts for class $c = 1,...,C$, is q_c, the total number of counts for all classes is $\sum_{c=1}^{C} q_c$. Suppose further the total number of voxels of all sets is M, then one has $\Pr(c \mid \mathbf{x}') = q_c / M$. The probability of a voxel mapped to $\mathbf{x}'$, disregarded of its class attribute, is $\Pr(\mathbf{x}') = \sum_{c=1}^{C} q_c / M$. The relative weight of class $c = 1,...,C$, is $t_c = q_c / \sum_{i=1}^{C} q_i$. Then $\hat{\mathbf{G}} = \sum_{c=1}^{C} \mathbf{G}_c t_c$ represents the possible class attribute for a specific voxel without considering the relative probability distribution.

3.3.3 Constructed probabilistic atlas

We have chosen 23 sets of abdominal CT data with different numbers of slices with a typical 40 slices. The slice spacing is 5mm and each slice is digitized into 512-by-512 pixels. The physical pixel intervals within a slice range from 0.5mm to 0.75mm for different sets. These sets of data are first scaled and translated to form isotropic data and all warping transforms are based on such isotropic data sets. We consider six organs, i.e., liver, right/left kidney, spleen, abdominal aorta and spinal canal. As a preliminary attempt, only four pairs of landmarks are used for all warping transforms and in the optimization. In Fig. 7, the left illustrates the quantity $\hat{\mathbf{G}}$ and the right is the probabilistic atlas $\mathbf{G}$ for the slice. The organs can be observed from the resulted atlas. However, the quality of the constructed atlas is relatively worse than a previous work[11] as we only use four landmarks instead of 36-point TPS.

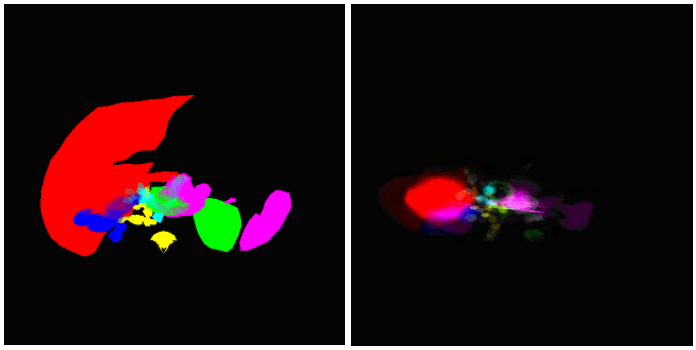

Figure 7. Two abdominal probabilistic atlases constructed using four pairs of landmarks

4. PERFORMANCE BENCHMARKING OF LIVER SEGMENTATION ALGORITHMS

Performance benchmarking of liver segmentation algorithms was performed in this study using 10 data sets selected from the constructed abdominal CT image DB and the corresponding ground truths. 2D contour atlases were constructed from data set No. 7 presented in Table 1. Two quantitative measure, sensitivity (SEN) and Jaccard similarity measure (JSM)[16] were calculated to assess the similarity between the computed and manually defined liver areas:

$$SEN = \frac{\#TPs}{\#GT} \quad \text{and} \quad JSM = \frac{\#TPs}{\#GT + \#FPs}, \tag{4}$$

where # is the cardinality of a set, TPs are true positives, FPs are false positives and GT is ground truth. Higher SEN value means that results contain more TPs and higher JSM indicates that results contain more TPs or with less FPs . We have experimented two automatic liver segmentation algorithms: the active contour method[9] and the level-set method[17]. The first one is a revised version as reported in reference 9, which is based on single 2D atlas. The second one is based on a variational level set[17] while using the composite 2D contour atlas in initialization. In both methods we make use of 2D atlases as the initialization contours.

4.1 Algorithm 1: Multiple 2D Atlases-Based Active Contours

An automatic liver segmentation algorithm which originates from a previously developed active contour with one single 2D atlas[9] was tested on the collected data set. Multiple 2D atlases are used to accommodate the large liver shape changes from the top to the bottom sections. The original algorithm consists of three stages. Firstly, a global transformation combining outer body contour extraction and iterative closest point algorithm is conducted to register the atlas to the target image. Then a local transformation is applied iteratively to bring the atlas contour of each organ closer to the target contour, followed by an active contour model (snake) with gradient vector flow to refine the segmentation. For the revised algorithm, multiple 2D atlases instead of one single atlas are used and its implementation is as follows: Firstly, manual segmentation results of three slices in one data set (No. 7 in Table 1 below) were chosen as 2D atlases. Note that these atlases should be able to indicate the shape changes of the liver from top to bottom sections. Then after the global and local transformations, a 2D atlas-based snake algorithm is performed on each data set and every slice in the data set is segmented by the initialization from one of the atlases. In this procedure, which atlas is chosen depends on the difference between the atlas and the target images. If the difference is less, the result would be better. This makes the method semi-automatic in initialization for the whole set.

4.2 Algorithm 2: Variational Level-Set Method Based on a Composite 2D Atlas

Level set based methods have been popular for medical image segmentation in recent years. In comparison with the active contour method level set based methods have many advantages in handling objects with complex shapes. In our experiments we make use of the composite 2D atlas as the initialization contour. Therefore only one 2D atlas is used to segment all slides, which makes the liver segmentation fully automatic after the initialization.

In traditional level set methods, the level set function can develop shocks, very sharp shape during the evolution, which makes further computation highly inaccurate. A common numerical scheme to avoid these problems is to initialize the level set function as a signed distance function before the evolution, and then re-initialize the level set function as the signed distance function periodically during the evolution. The re-initialization process is very complicated and expensive. Here we implemented a variational level set method for the detection of liver boundaries proposed in.[17] It can force the level set function to be close to the signed distance function, and thus eliminates the need of the re-initialization procedure. This formulation consists of an internal energy term which penalizes the deviation of the level set function from a signed distance function and an external energy which drives the motion of the zero level set towards the object boundaries. The resulted the level set function is the gradient flow that minimizes the overall energy.

5. RESULTS AND DISCUSSION

5.1 Benchmarking Results

We extract all the 2D atlases from the data set No.7 shown in Table 1. Figs. 8 and 11 show the segmentation results of the two algorithms with segmented contours are overlaid on corresponding CT images. For some slices, the algorithm can find satisfying liver boundaries. However when certain situations are met such as low gradient to neighboring organs (slice 7), two unconnected component in 2D presentation (slice 7) and large shape changes (slice 19), its performance is poor. The *SEN* and *JSM* curves over each sequential slice in this set are shown in Fig. 9. It can be observed that major segmentation errors occur at the sections where liver has in connection with heart, liver appears to be two parts without connection in 2D and liver shapes have large change among neighboring slices.

In experiments using the level-set method, the same set is used as the sample to generate a universal initial contour. In the beginning, three images (slice 7, 13 and 30) are chosen from the set to construct a multi-2D atlas (described above). This contour of the atlas is then used to initialize level sets in segmentation experiments for all other data sets. Similarly, Fig.11 shows selected segmentation results from the same data set as that for the previous method. The level set method can find more than one objects and shrink to multiple close contours as seen from the result for slice 3 where one region in the right is wrongly segmented. We use all regions enclosed in the possibly multiple contour(s) as the segmented liver and evaluate the algorithm performance.

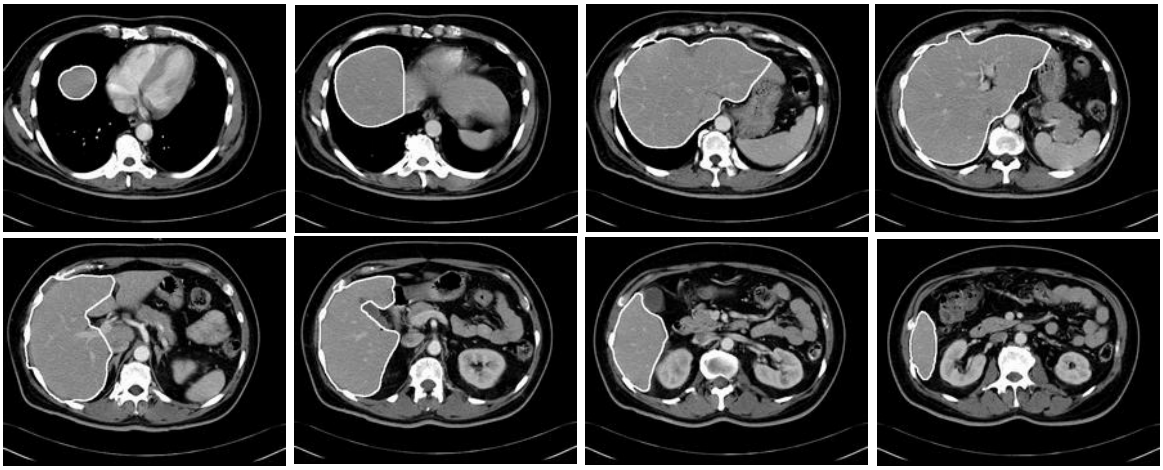

Figure 8. Selected segmentation results for Algorithm 1 for data set 7.

From left, top row: slice 3,7,11,15; bottom: slice 19,23,27,31

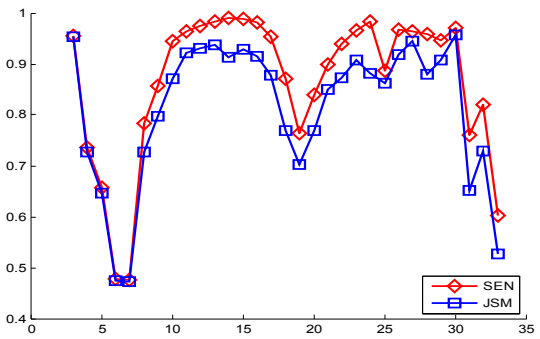

Figure 9. SEN and JSM curves over each sequential slice data set 7 for Algorithm 1

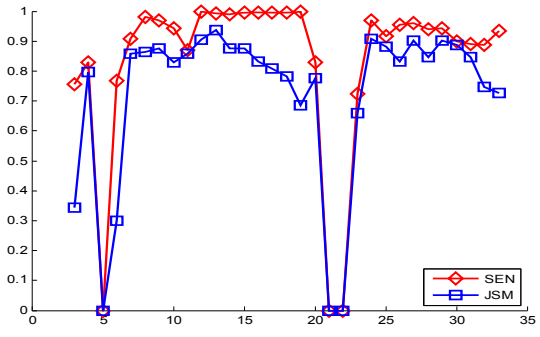

Figure 10. SEN and JSM curves over each sequential slice in data set 7 for Algorithm 2

The *SEN* and *JSM* curves over each sequential slice in this set are shown in Fig. 10. The segmentation results are not very good for some slices. For example, for slice 7, it also includes the heart part. For some slices, there are no overlap regions and hence yield zeros for *SEN* and *JSM* . This may be due to wrong choices of objects since the largest object found may not be the liver. However, we also find that, for some slices, it can perform pretty well, like the previous method. Table 1 shows the quantitative measures for performance benchmarking of liver segmentation algorithms for the same 10 data sets. The level-set method normally has a higher sensitivity while a lower *JSM* thus including more false positive segmentation.

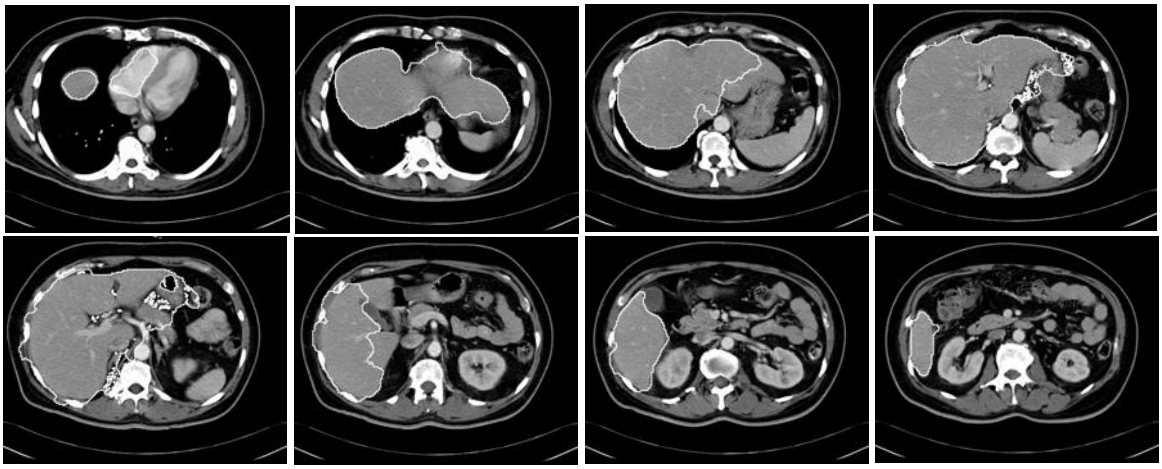

Figure 11. Selected segmentation results using a level-set method for data set 7:

From left, top row: slice 3,7,11,15; bottom: slice 19,23,27,31

Table 1 Quantitative measures for performance benchmarking of liver segmentation algorithms

	Data No.	1	2	3	4	5	6	7	8	9	10	**Ave.**
A. 1	***SEN***	0.83	0.94	0.86	0.94	0.85	0.81	0.88	0.88	0.74	0.78	0.85
	JSM	0.71	0.79	0.68	0.78	0.71	0.75	0.83	0.81	0.65	0.69	0.74
A. 2	***SEN***	0.95	0.97	0.96	0.96	0.95	0.93	0.93	0.91	0.93	0.95	0.94
	JSM	0.38	0.44	0.43	0.34	0.48	0.37	0.73	0.41	0.32	0.46	0.43

5.2 Discussion

Preliminary experimental results from the two algorithms which make use of the 2D contour atlas have shown comparable performances in liver segmentation as reported in the recent international medical image segmentation workshop held in MICCAI07[10]. So far we have only used 2D contour atlases in our experiments, much more work to be done in the future in which various types of knowledge about the livers can be extracted from various levels, including voxels, curves and surfaces, and incorporate the extracted knowledge into liver segmentation algorithm to make segmentation more robust.

Results using single atlas and an active contour method can be found in reference.[9] We have also experimented in this way and found the same conclusion: for slices near the slice where the atlas is taken, the segmentation results are much better than those far away. The contour of the atlas we used to initialize the level-set algorithm is shown in Fig. 5. This motivates us using more atlases. In the above experiments, we have tested two ways to use three different atlases.

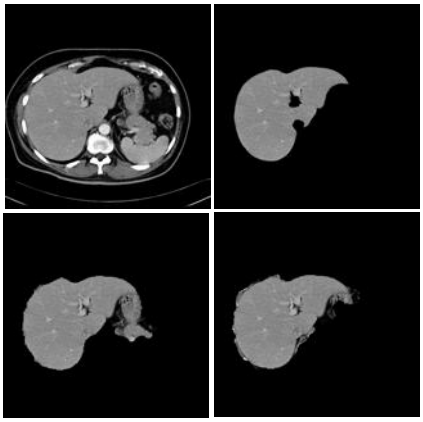

Figure 12. Applying the level-set method, using a single composite 2D atlas produces better results than using a single 2D atlas: From left, original image, ground truth of liver, result using a single 2D atlas and that using a composite 2D atlas.

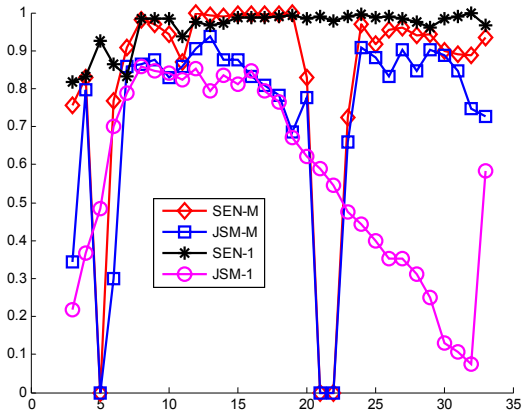

Figure 13. Performance comparison: level-set method using single 2D atlas and using single composite atlas generated from data set 7

In the experiments applying the active contours, the three atlases are extracted from slices in the beginning, the middle and the end positions, respectively. Each atlas covers separately about 10 slices around the position where it was extracted. In the experiments applying the level-set method, the three slices are chosen in the same way as before however used differently. A single composite 2D atlas is constructed and covers all slices in the scan set. The later method has higher sensitivity but significantly lower similarity due to obvious over segmentation, as shown in Table 1. The logic OR operation includes more region which is irrelevant to liver for most of slices for the initialization of evolution of level sets. Since the intensities and gradient properties of the liver and neighboring organs, such as the heart and the stomach, are very similar, the level-set method can not converge properly to the correct liver region resulting in many false alarms. The active contour method, on the other hand, has lower sensitivity.

Since we extract all the 2D atlases from the data set No.7 shown in Table 1, the similarity measure of this set is much higher than others for either method. Studying Figs. 9, 10 and 13, one can also observe that, the performance for the slices used to construct the atlas is better than those not-used slices. These observations remind us of the representativeness of the constructed atlas. If the atlas does not represent certain types of abdominal data, then the segmentation results will have low similarity.

To compare the performance using a single composite 2D atlas and that using single atlas, we have applied them in two further experiments using the level set method. We can achieve better results using a multi-2D atlas (shown in Fig, 6) than using single atlas, as illustrated in Fig. 12. Comparing with the ground truth presented in the same figure, the single atlas method includes a part of the stomach. Such a false inclusion may be due to a wrong atlas used. Fig. 13 shows the sensitivity and the similarity measures of the level-set method applying both methods on the same scan set of 34 slices, where "SEN-M" and "JSM-M" are sensitivity and similarity measures for the multi-2D atlas and "SEN-1" and "JSM-1" are for the single atlas method. On average over the whole set, the single-atlas method has sensitivity 0.965 and similarity 0.583, respectively, whereas the multi-2D-atlas method achieves sensitivity 0.934 and a higher similarity 0.728, respectively. The later method captures more true positive segmentation.

The definitions of liver volume and performance measure protocols are also important issues. In clinical practice for the calculation of effective liver volume, major vascular structures with sinuses inside the liver such as artery and portal vein are excluded from liver volume. This protocol was adopted in our organ ground truth tracing procedure. While for edge detection-based methods such as active contour and level set, these internal hollow structures cannot be detected and the whole outline with the enclosed area was considered as the resultant target object. It can partially explain the large false positives we met in our experiments. In some other investigations, scans without contrast enhancement or at delayed phase were performed, therefore due to the weak difference in densities between liver tissue and vessels, liver ground truth was traced without excluding major vascular structures and sinuses. Under this definition for liver volume, some methods may achieve better performance. For the liver segmentation contest workshop in MICCAI 2007, liver contour ground truth was defined as the liver outer boundary without the exclusion of major vascular structures and sinuses. [10] Although overlap ratio (sensitivity) is the most common indicator to measure the segmentation performance,

research efforts have been devoted in this filed and a set of complicated performance measures, including volumetric overlap, relative absolute volume difference, average symmetric absolute surface distance, symmetric RMS surface distance and maximum symmetric absolute surface distance, have been proposed for 3D object segmentation.[18] In the present study, considering that we still worked on a 2D slice by slice basis, sensitivity and JSM were used as the evaluation measures.

The purpose of construction of a probabilistic atlas is to automate the segmentation and volume quantification of abdominal organs including the liver. We plan to further explore this aspect.

Finally, we believe in order to achieve robust segmentation performance it is essential to make use of knowledge extracted from collections of liver image databases and associated ground truth databases; the knowledge can be represented at many levels such as pixel/voxel density's statistical distributions, single or multiple statistical 2D contour atlases and statistical 3D surface atlases for liver, and multi-organ probabilistic atlases for abdominal organs. These atlases can be constructed for both normal and pathological livers, depending applications' needs of liver segmentations, such as liver transplantation or liver tumor detection for radiation planning, etc.

6. CONCLUSION AND FUTURE WORK

We have built up a generic platform for benchmarking liver image segmentation algorithms and volume estimation. It contains an abdominal CT image DB, a ground truth DB, some annotation tools, performance measure protocols, and preliminary statistical models such as 2D/3D atlases for individual and multiple organs. We have also benchmarked two liver segmentation methods, with preliminary results. The experimental results have shown that, using multiple 2D contour atlases can achieve significantly better results than using single atlas. We plan to improve the atlas models and extract more knowledge from liver image databases and other ground truth databases and incorporate them into liver segmentation algorithms. We also plan to develop more integrated tools to facilitate the benchmarking. We target to increase the CT scans to about 300 sets in the near future and plan to make the DBs built available to medical imaging research community for performance benchmarking of liver segmentation algorithms.

ACKNOWLEDGEMENT

The research work is supported by the Singapore Bioimaging Consortium (SBIC), Agency for Science Technology and Research (A*STAR), with grant SBIC RP C-008/2006. We thank Feng Ding who provides segmentation codes using an active contour method and Fengshou Yin who helps on segmentation using a level-set method.

REFERENCES

1. K. Doi and H.K. Huang, "Computer-aided diagnosis (CAD) and image guided decision support," Computerized Medical Imaging and Graphics, **31(4-5)**, 195-197 (2007).
2. K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," Computerized Medical Imaging and Graphics, **31(4-5)**, 198-211 (2007).
3. S.G. Armato III, G. McLennan, M.F. McNitt-Gray, et al., "Lung Image Database Consortium: Developing a resource for the medical imaging research community," Radiology, **232(3)**, 739-748 (2004).
4. M.W. Vannier, E.V. Staab and L.C. Clarke, "Medical image archives – present and future", Proceedings of the International Conference on Computer-Assisted Radiology and Surgery (CARS), 2002.
5. W. Hersh and H. Müller, "Image retrieval: Image retrieval in medicine: The ImageCLEF medical image retrieval evaluation", Bulletin of the American Society for Information Science & Technology, February /March 2007.
6. L. Gao, D.G. Heath, Kuszyk BS, et al., "Automatic liver segmentation technique for three-dimensional visualization of CT data," Radiology, **201(2)**, 359-364 (1996).
7. Y. Masutani, K. Uozumi, M. Akahane, et al., "Liver CT image processing: a short introduction of the technical elements," Europe Journal of Radiology, **58(2)**, 246-251 (2006).
8. J. Lee, N. Kim, H. Lee, et al., "An automatic method for fast and accurate liver segmentation in CT images using a shape detection level set method," Proceedings of SPIE Medical Imaging, vol. 6512, 2007.

9. F. Ding, W.K. Leow and S.C. Wang, "Segmentation of 3D volume images using a single 2D atlas," Lecture Notes in Computer Science, vol. 3765, 459-468 (2005).
10. T. Heimann, B. van Ginneken and M. Styner, Workshop on 3D Segmentation in the Clinic: A Grand Challenge, http://mbi.dkfz-heidelberg.de/grand-challenge2007/sites/orga.htm
11. H. Park, P.H. Bland and C.R. Meyer, "Construction of an abdominal probabilistic atlas and its application in segmentation," IEEE Transactions on Medical Imaging, **22(4)**, 483-492 (2003).
12. C.R. Meyer, J.L. Boes, B. Kim, 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," Medical Image Analysis, **1(3)**, 195-206 (1997).
13. F.L. Bookstein, "Shape and the information in medical images: a decade of the morphometric synthesis," Computer Vision and Image Understanding, **66(2)**, 97-118 (1997).
14. W.T. Vetterling, W.H. Press, S.A. Teukolsky, et al., *Numerical Recipes in C: The Art of Scientific Computing*, Second Edition, Cambridge University Press, Cambridge, U.K., 1992.
15. M.M. Cerney, D.C. Adams, and Judy M. Vance, "Image Warping of Three-Dimensional Body Scan Data," Proceedings of the SAE Digital Human Modeling Conference, vol. 1, 2003.
16. Z.Y. Shan, Q. Ji, A. Gajjar, et al., "A knowledge-guided active contour method of segmentation of cerebella on MR images of pediatric patients with medulloblastoma," Journal Magnetic Resonance Imaging, **21(1)**, 1-11 (2005).
17. C. Li, C. Xu, C. Gui, et al., "Level Set Evolution Without Re-initialization: A New Variational Formulation," Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 430-436 (2005).
18. M. Gerig and M. Chakos, "Valmet: a new validation tool for assessing and improving 3D object segmentation," Proceedings of the International Conference on Medical Imaging Computing and Computer Assisted Intervention (MICCAI), 516-523 (2001).