Medical Physics
The International Journal of Medical Physics Research and Practice

Multiatlas whole heart segmentation of CT data using conditional entropy for atlas ranking and selection

Xiahai Zhuang, Wenjia Bai, Jingjing Song, Songhua Zhan, Xiaohua Qian, Wenzhe Shi, Yanyun Lian, and Daniel Rueckert

Citation: Medical Physics **42**, 3822 (2015); doi: 10.1118/1.4921366
View online: http://dx.doi.org/10.1118/1.4921366
View Table of Contents: http://scitation.aip.org/content/aapm/journal/medphys/42/7?ver=pdfcov
Published by the American Association of Physicists in Medicine

Multiatlas whole heart segmentation of CT data using conditional entropy for atlas ranking and selection

Xiahai Zhuang[a)]
SJTU-CU International Cooperative Research Center, Department of Engineering Mechanics, School of Naval Architecture Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Wenjia Bai
Biomedical Image Analysis Group, Department of Computing, Imperial College London, 180 Queens Gate, London SW7 2AZ, United Kingdom

Jingjing Song and Songhua Zhan
Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

Xiaohua Qian
SJTU-CU International Cooperative Research Center, Department of Engineering Mechanics, School of Naval Architecture Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Wenzhe Shi
Biomedical Image Analysis Group, Department of Computing, Imperial College London, 180 Queens Gate, London SW7 2AZ, United Kingdom

Yanyun Lian
Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China

Daniel Rueckert
Biomedical Image Analysis Group, Department of Computing, Imperial College London, 180 Queens Gate, London SW7 2AZ, United Kingdom

(Received 22 October 2014; revised 5 May 2015; accepted for publication 8 May 2015; published 9 June 2015)

Purpose: Cardiac computed tomography (CT) is widely used in clinical diagnosis of cardiovascular diseases. Whole heart segmentation (WHS) plays a vital role in developing new clinical applications of cardiac CT. However, the shape and appearance of the heart can vary greatly across different scans, making the automatic segmentation particularly challenging. The objective of this work is to develop and evaluate a multiatlas segmentation (MAS) scheme using a new atlas ranking and selection algorithm for automatic WHS of CT data. Research on different MAS strategies and their influence on WHS performance are limited. This work provides a detailed comparison study evaluating the impacts of label fusion, atlas ranking, and sizes of the atlas database on the segmentation performance.
Methods: Atlases in a database were registered to the target image using a hierarchical registration scheme specifically designed for cardiac images. A subset of the atlases were selected for label fusion, according to the authors' proposed atlas ranking criterion which evaluated the performance of each atlas by computing the conditional entropy of the target image given the propagated atlas labeling. Joint label fusion was used to combine multiple label estimates to obtain the final segmentation. The authors used 30 clinical cardiac CT angiography (CTA) images to evaluate the proposed MAS scheme and to investigate different segmentation strategies.
Results: The mean WHS Dice score of the proposed MAS method was 0.918 ± 0.021, and the mean runtime for one case was 13.2 min on a workstation. This MAS scheme using joint label fusion generated significantly better Dice scores than the other label fusion strategies, including majority voting (0.901 ± 0.276, $p < 0.01$), locally weighted voting (0.905 ± 0.0247, $p < 0.01$), and probabilistic patch-based fusion (0.909 ± 0.0249, $p < 0.01$). In the atlas ranking study, the proposed criterion based on conditional entropy yielded a performance curve with higher WHS Dice scores compared to the conventional schemes ($p < 0.03$). In the atlas database study, the authors showed that the MAS using larger atlas databases generated better performance curves than the MAS using smaller ones, indicating larger atlas databases could produce more accurate segmentation.
Conclusions: The authors have developed a new MAS framework for automatic WHS of CTA and investigated alternative implementations of MAS. With the proposed atlas ranking algorithm and joint label fusion, the MAS scheme is able to generate accurate segmentation within practically acceptable computation time. This method can be useful for the development of new

clinical applications of cardiac CT. © *2015 American Association of Physicists in Medicine.* [http://dx.doi.org/10.1118/1.4921366]

Key words: whole heart segmentation, multiatlas segmentation, atlas ranking, label fusion, cardiac CT

1. INTRODUCTION

According to the World Health Organization, cardiovascular diseases are the number one cause of deaths globally. In 2008, 17.3×10^6 people died from cardiovascular diseases, accounting for 30% of deaths around the world.[1] Medical imaging and image analysis have had significant impacts on healthcare. Cardiac computed tomography (CT), in particular, the contrast enhanced CT and CT angiography (CTA), can capture the anatomy of the heart and it has been widely used in the clinic.[2] The morphological and pathological information from three-dimensional (3D) medical images is useful in diagnosis and treatment of patients. To enable the development of novel clinical applications and thus improve cardiology, accurate extraction and precise interpretation of the anatomical information become particularly important.

The goal of whole heart segmentation (WHS) is to extract the substructures of the heart, including the four chamber blood cavities and the left ventricular (LV) myocardium. In some cases, the great vessels are also of interest.[3] Figure 1 shows an example of a cardiac CT image and its corresponding WHS result. Notice that WHS is different from other applications of cardiac segmentation which focuses solely on the ventricles and/or myocardium,[4] or the isolation of the whole heart volume.[5]

WHS is useful for a number of clinical applications. For example, it can be directly used to extract functional indices such as the ejection fraction and myocardial mass. It is also expected that the functional analysis of the whole heart has the potential of detecting subtle functional abnormalities or changes of the heart.[6] This is important for the diagnosis of patients who otherwise have normal systolic function of the ventricles but are suspected to have abnormal function in other regions. Furthermore, the 3D surface rendering, as shown in Fig. 1(c), has a wide range of applications. For example, it can be used to investigate congenital malformations of the heart, or to guide interventional procedures by fusing the 3D surface with a real-time fluoroscopy ultrasound scan.

1.A. Challenges and related works

Although extracting the whole heart structure is essential, manual delineation of all the substructures is labor-intensive and tedious, and the results can be subject to intra- and interobserver errors. Therefore, automating the segmentation becomes increasingly important. However, achieving fully automatic WHS from cardiac CT is arduous. The challenge mainly comes from (1) the large shape variations of the cardiac anatomy, (2) the inconsistent and variable texture patterns of contrast enhanced CT images, and (3) the indistinct boundaries of substructures in the images.[3]

To achieve automatic WHS, *a priori* models are commonly used to guide the segmentation.[3] Lotjonen *et al.*[7] and Koikkalainen *et al.*[8] employed statistical shape models to estimate the shape of the heart in cardiac magnetic resonance (MR) images. Zheng *et al.*[9] hierarchically detected boundary landmarks based on steerable features and marginal space learning. The shape was regularized by projecting the deformed model onto a subspace of a learned statistical shape model. Peters *et al.*[10] developed a deformable model-based method for the WHS of both cardiac CT and MR images. The same paper proposed a piecewise affine parametric adaptation method to extend the shape variability of the model and employed the simulated search for detecting the optimal response of edges. Kirisli *et al.*[11] used eight atlases to perform a multicenter, multivendor evaluation study on the WHS of CT data. Their study adopted a conventional scheme, based on a global affine registration and a free-form deformation (FFD) registration, for atlas-to-target image registration and the majority voting method was used for label fusion. Zhuang *et al.*[12] employed a mean atlas and a comprehensive registration algorithm for the atlas-based WHS of MR images. More specifically, the locally affine registration method (LARM) was proposed to tackle the large shape variability of the cardiac anatomy. By adopting LARM, the atlas-based segmentation method yielded a mean WHS Dice score of 0.84 ± 0.05 in a challenging test data set involving nine different pathologies.

(a) (b) (c)

FIG. 1. Demonstration of an atlas: (a) CT data as atlas intensity image; (b) manual segmentation as atlas label image; (c) volume rendering of whole heart substructures.

Atlas-based segmentation has gained great popularity in tackling challenging segmentation tasks. An atlas is defined as a pair of intensity and label images, and a segmentation is obtained via atlas-to-target registration and subsequent segmentation propagation. Multiatlas segmentation (MAS) involves fusing multiple classifiers to derive a single segmentation. This is a useful strategy to improve the robustness and accuracy of atlas-based segmentation.[11,13–22] It is particularly effective and efficient when there are multiple images with manual labeling available as atlases such as in brain MR segmentation[15–20] and cardiac image segmentation.[11,17,21]

Rohlfing *et al.*[13] and Rohlfing and Maurer[14] studied different atlas ranking and selection methods for MAS and showed that the intensity-based similarity metrics between the warped atlas image and the target image were good criteria for atlas selection. Heckemann *et al.*[16] proposed a mathematical model for predicting the convergence of segmentation accuracy with respect to the increased number of atlases selected for label fusion. Aljabar *et al.*[15] and van Rikxoort *et al.*[17] showed that with a good ranking criterion, the accuracy of MAS could reach a global optimum after fusing a certain number of selected atlases. Intensity-based similarity measures, such as the mutual information (MI) or normalized mutual information (NMI),[23–25] are often used for atlas ranking and selection.

However, the intensity-based similarity measures do not necessarily represent the segmentation quality of each atlas. This is particularly evident in cardiac CTA where the texture patterns can vary greatly. Figure 2 provides an example for illustration:

— (a) and (b) are the intensity images of two atlases, respectively, referred to as atlas-A and atlas-B;
— (c) and (d) are, respectively, the target image and the image superimposed with manually delineated contours;
— (e) and (f) show the target image superimposed with the contours of two segmentation results propagated from atlas-A and atlas-B, respectively;
— (g) and (h) are, respectively, the deformed intensity images of atlas-A and atlas-B after atlas-to-target registration.

For atlas-A, the segmentation accuracy is 0.831 (Dice score), and the NMI between the deformed atlas image and the target image is 1.113; for atlas-B, the segmentation accuracy is 0.802, and the NMI is 1.132. In terms of segmentation accuracy, atlas-A performs better than atlas-B. However, atlas-B has a higher similarity value (NMI), thus can be erroneously selected as a better choice over atlas-A if the NMI-based similarity measure is used as the criterion for atlas ranking and selection.

In this work, we propose a new atlas ranking method based on conditional entropy. This method computes the negative conditional entropy of the target image given the propagated atlas labeling, which incorporates the segmentation performance of each atlas and is expected to perform better than the conventional intensity-based approaches.

In MAS, the majority voting algorithm has been widely used to fuse multiple classifiers.[26] Recently, a number of approaches have been proposed to improve the performance of label fusion by using locally weighted voting and patch-search strategies. In these approaches, the vote is spatially weighted according to the intensity similarity between the atlas and the target image and a patch search is employed to recover local misalignments.[16,18,21,27–36] Bai *et al.*[21] proposed a probabilistic patch-based fusion (PPF) method and formulated the label fusion problem in a Bayesian framework. This model considers not only the similarity between local patches but also it accounts for the registration uncertainty between the target image and the atlas. A common limitation of label fusion

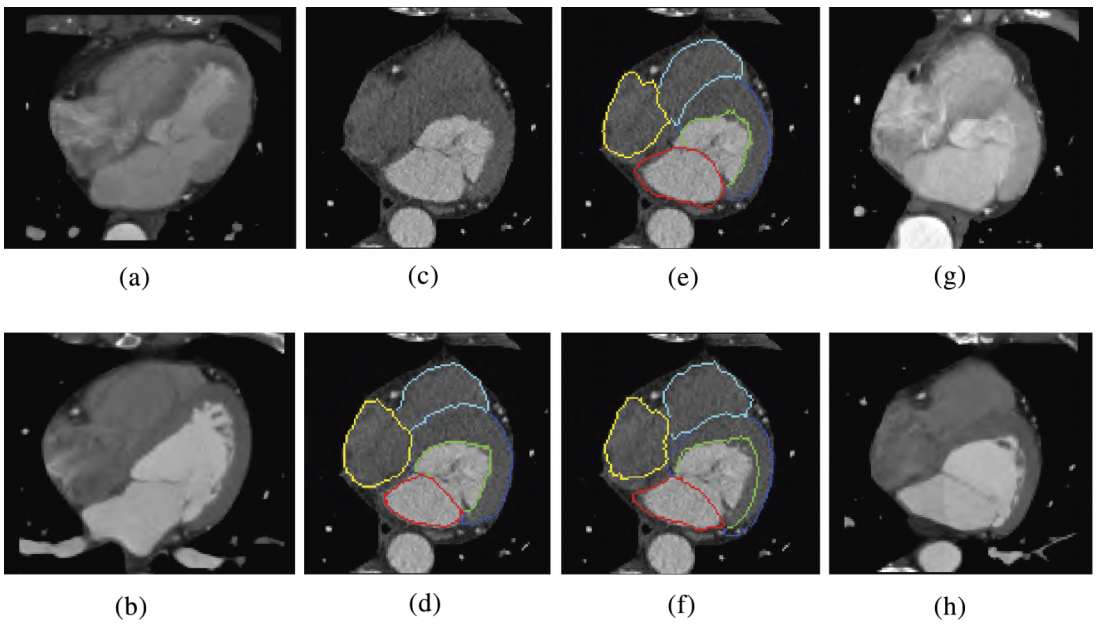


FIG. 2. Illustration of the atlas ranking problem using the intensity-based similarity measures. Please refer to the text for detail.

methods is that the weights are computed independently for each atlas, without taking into account the fact that different atlases may produce similar labeling errors. Wang *et al.*[35] proposed a joint label fusion (JLF) algorithm to estimate local weights by minimizing the expectation of labeling errors. This algorithm explicitly models the pairwise correlation between atlases. The joint probability is estimated using patch-based intensity correlation. Patch-based strategies have been shown to be effective in improving performance of label fusion, but the impacts of patch size and search range have not been fully investigated.

1.B. Contributions of this work

The main contributions of this work are as follows:

First, we propose a novel atlas ranking metric, which incorporates the propagated atlas labeling, to improve the MAS scheme.

Second, we demonstrate a clinically applicable MAS scheme which can generate robust and accurate WHS of CT data. This MAS scheme adopts the proposed atlas ranking algorithm and a hierarchical image registration method specifically designed for WHS of cardiac images. In our previous work, we have demonstrated that the three-level LARM can improve the atlas-to-target registration of cardiac images and consequently improve the segmentation accuracy of the single atlas-based approach.[12] In this work, we further investigate the performance of this registration technique for MAS. Details of the WHS results are reported to provide a benchmark for future research.

Finally, we provide detailed studies investigating different MAS strategies in order to understand the impacts of different label fusion algorithms, atlas ranking methods, and atlas databases. While it has been widely recognized that the MAS should outperform the single atlas-based segmentation (SAS), the influence of the sizes of atlas databases on the MAS performance has not be well understood. In this work, we employed a large number of cardiac CTA images and performed a number of experiments to study this aspect.

The remainder of the paper is organized as follows: Sec. 2 introduces the data and atlases used in this work and Sec. 3 describes the methods. The results are presented in Secs. 4 and 5 provides a discussion of these results. Finally, we conclude the work in Sec. 6.

2. MATERIALS

2.A. Data

Thirty cardiac CTA images from 30 patients at the end-diastolic phase were used. All patients were Chinese, aged from 43 to 86 (64.8 ± 11.6) yr old, with body weight between 51 and 93 (69.0 ± 9.37) kg and height between 153 and 178 (164.4 ± 7.8) cm. Among them, 13 were male and 17 were female. All subjects had cardiovascular diseases, involving a wide variety of pathologies. Some patients had a combination of several different types of pathologies. For details of the demographics and the pathologies, please refer to Table I.

The CTA data were acquired from state-of-the-art 64-slice CT scanners (Philips Medical Systems, the Netherlands) using a standard coronary CTA protocol at two sites. All the data cover the whole heart from the upper abdominal to the aortic arch. Each slice, acquired in the axial view, has a wide field-of-view (FOV) as the example shown in Fig. 1(a). The in-plane resolution is about 0.44×0.44 mm and the average slice thickness is 0.60 mm. The orientation information was recorded in the header of the original files, which were acquired and stored in the Digital Imaging and Communications in Medicine (DICOM) format. This information was utilized in the atlas-to-target registration.

2.B. Substructures of interest and manual segmentation

Seven substructures were of interest in the WHS study, including

(1) the LV cavity;
(2) the right ventricular (RV) cavity;
(3) the left atrial (LA) cavity;
(4) the right atrial (RA) cavity;
(5) the myocardium of the left ventricle (Myo) where the epicardium (Epi) is assessed in the evaluation of surface delineation;
(6) the ascending aorta trunk from the aortic valve to the superior level of the atria;
(7) the pulmonary artery (PA) trunk from the pulmonary valve to the bifurcation point.

The substructures in each image were manually labeled to generate the atlas label map and to provide a gold stan-

TABLE I. Summary of the subjects and CTA data used.

Resolution	~0.44×0.44 mm in-plane and ~0.60 mm slice thickness		
Subjects	Number: 30 Ethnicity: Chinese		
Gender	All (30 subjects)	Male (13 subjects)	Female (17 subjects)
Age (yr)	43 – 87 (64.8 ± 11.6)	43 – 84 (59.9 ± 11.4)	54 – 87 (68.5 ± 10.7)
Weight (kg)	51 – 93 (69.0 ± 9.37)	58 – 91 (73.8 ± 13.6)	51 – 93 (65.5 ± 10.8)
Height (cm)	153 – 178 (164.4 ± 7.8)	170 – 178 (173.0 ± 3.9)	153 – 165 (158.4 ± 3.8)
Pathologies	(1) Cardiac function insufficiency (NYHA II), (2) cardiac edema, (3) hypertension (III), (4) sick sinus syndrome, (5) arrhythmia, (6) atrial flutter, (7) atrial fibrillation, (8) artery plaque, (9) coronary atherosclerosis		

dard for validation as well. Figures 1(a) and 1(b) show an example of a cardiac CT image and its corresponding manual segmentation. The manual segmentation was performed by two well-trained master students majoring in biomedical engineering, and neither of them was aware of the methodologies of automatic segmentation. The two observers employed a brush tool in the software ITK-SNAP (Ref. 37) in order to manually label each substructure slice-by-slice. The labeling results were double checked by an expert in cardiac anatomy. The manual segmentation took 6–10 h/case.

2.C. Atlas

The cardiac CT images were used as intensity images for the atlas. Figure 1(a) shows an example of the intensity image of a cardiac CT atlas. The atlas label image indicates the segmentation of all substructures of interest using a set of discrete labels. Figure 1(b) shows an example of the label image of a cardiac CT atlas.

3. METHODS

This section explains the different methodologies used, including the SAS and the MAS (Sec. 3.A), the atlas-to-target image registration (Sec. 3.A.1), the atlas ranking and selection techniques (Sec. 3.A.2), the label fusion strategies (Sec. 3.A.3), the evaluation measures (Sec. 3.B), and the experiments designed for the proposed studies (Sec. 3.C).

3.A. Single atlas-based segmentation and multiatlas segmentation strategy

Figure 3(a) shows the conventional atlas-based segmentation, i.e., the SAS scheme, where an atlas is registered to the target image and the labeling of the atlas is then propagated to the target image by applying the resulting transformation to the atlas label image. SAS may generate a poor result if the registration algorithm does not perform well. In contrast, MAS accounts for intersubject variation of the anatomy and texture patterns by using a set of atlases. It reduces the risk of poor segmentation from a single registration by combining multiple results from the set of atlases and has shown great potentials in segmenting complex organs.[13–16]

Let I be the target image; given a set of N atlases $\{(A_n,L_n)\}$, $n = 1,\ldots,N$, where A_n and L_n are a pair of atlas intensity image and atlas label image. MAS performs an atlas-to-target registration for each atlas,

$$T_n = \underset{T_n}{\text{argopt}}\ \text{Registration}(I,A_n),$$

where T_n is the resulting transformation. The segmentation of I by atlas A_n, denoted as S_n, is then derived by $S_n = T_n(L_n)$, which is the warped atlas label image by T_n. The MAS scheme, as illustrated in Fig. 3(b), selects a subset of M atlases based on an atlas ranking criterion (Sec. 3.A.2), and the final result is derived by a label fusion step,

$$S = \text{Label Fusion}(\{S_1,S_2,\ldots,S_M\}),$$

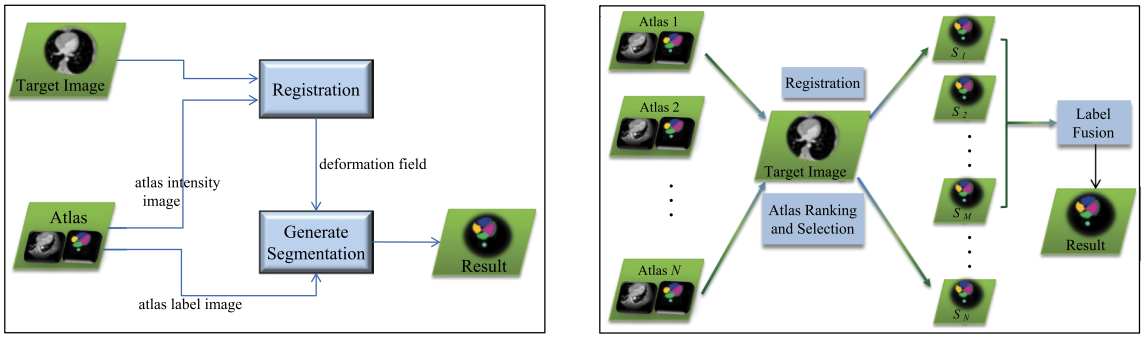
where $\{S_i\}$ are the deformed atlas label images from the selected M atlases.

3.A.1. Atlas-to-target registration

Since the atlas and target images generally come from different subjects, deformable registration is often adopted to accurately align them. The conventional strategy employs a global affine registration, followed by a fully nonrigid algorithm.[11,38] However, due to the complex shape and large shape variability of the heart, nonrigid registration may generate unrealistic deformation fields due to the poor initial alignment of substructures by a global affine registration.[12]

We adopted a hierarchical registration framework, including a global affine registration for heart localization, a LARM for substructure initialization, and a FFD registration for local refinement. Note that in the heart localization, we used the orientation information contained in the original DICOM files of the atlas and the target image for orientation correction before performing intensity-based registration.[25] This hierarchical framework was specifically designed for the registration of cardiac images.[12]

3.A.1.a. Substructure initialization. LARM has been proposed in order to tackle the variation coming from the substructures of the heart, by modeling each of them using a locally affine transformation.[12,39] The global deformation field of



(a) (b)

FIG. 3. Segmentation frameworks: (a) SAS and (b) MAS.

LARM is computed as follows:

$$T(x)=\begin{cases} G_i(x), & x\in V_i,\ i=1,\ldots,n \\ \sum_{i=1}^{n} w_i(x)G_i(x_i), & x\notin \bigcup_{i=1}^{n} V_i \end{cases}, \quad (1)$$

where $\{V_i\}$ is the set of predefined substructures, $\{G_i\}$ is the set of assigned locally affine transformations, and $w_i(x)$ is a normalized weighting factor related to the inverse distance between a point x and the region V_i. To obtain a folding-free deformation field, a regularization step is applied by monitoring the Jacobian matrix J_T, such that when the determinant of the matrix at any point drops below a threshold (0.5 as recommended in Ref. 12), then (1) a new image pair is generated by applying current deformation field to the original images; (2) the locally affine transformations are all reset to identity; and (3) a new optimization process is started. The final transformation of the registration is a concatenation of the series of transformations,

$$T=T_m\circ T_{m-1}\circ\cdots\circ T_1. \quad (2)$$

The composed T is a nonfolding deformation field, since all the $\{T_i\}$ are folding free.

We adopted a three-level LARM procedure which could improve the robustness of the atlas-to-target registration:[12]

- First, we used LARM with two locally affine transformations. The corresponding local regions were the descending aorta and the rest of the whole heart structure. We chose this setting because some of the CT images did not fully cover the aortic arch, and the descending aorta was separated from the rest of the whole heart.
- Then, we used LARM with three locally affine transformations by dividing the whole heart into three local regions: the descending aorta, the ventricular region, and the combination of atria and great vessels.
- Finally, we used LARM with seven locally affine transformations. The seven local regions were the two ventricles, the two atria, the pulmonary artery, the ascending aorta, and the descending aorta.

3.A.2. Atlas ranking and selection

A number of works have demonstrated the importance of ranking and selecting a subset of atlases, instead of using all for label fusion in MAS.[13,15,16,21,28,40] This is because some atlases can generate poor segmentation results which affect the consensus in fusing the multiple classifiers. Therefore, it is common to rank the atlases with respect to the expected performance and then select a subset of atlases with *better* performance. Since the heart shape is less determined by metainformation such as age or demographics and the pathological information is usually unknown before diagnosis, ranking atlases using image information is often adopted. The commonly used strategies compute a voxelwise and intensity-based similarity measure such as MI or NMI (Refs. 13 and 15) to assess the expected segmentation quality of an atlas. These ranking criteria assume that the similarity measure can indicate the expected segmentation quality, which may not always hold as Fig. 2 illustrates.

We propose a new metric, the negative conditional entropy of the target image given the propagated labeling (the warped label image S_n), as the atlas ranking criterion,

$$\mathcal{R}(A_n)\equiv -(H(I|S_n))=\sum_{i\in I,l\in S_n} p(i,l)\log\frac{p(i,l)}{p(l)}, \quad (3)$$

where H denotes entropy, $p(i,l)$ is the joint probability of the intensity value i and label l, and $p(l)$ is the marginal probability of label l.

Conditional entropy measures the uncertainty inherent in the intensity distributions of the target image, conditioned on the propagated atlas label image. Compared with the intensity-based similarity measures, conditional entropy provides a mechanism to compute the relationship between the target image and the atlas segmentation. Hence, it has the potential of providing less biased atlas ranking.

To study the influence of different atlas selection strategies on performance of MAS, we compared the proposed method, referred to as *Conditional entropy (nonrigid REG)*, with four other ranking methods:

(1) *Manual ranking (ground truth)*: This method computes the generalized Dice score[3,41] between the atlas segmentation and gold standard segmentation to rank the atlas. The computation is described in Eq. (5) (Sec. 3.B). Note that this method provides the best ranking result but is unavailable in practice, since the gold standard segmentation is unknown.
(2) *NMI (nonrigid REG)*: This method computes the NMI (Ref. 25) between the target image and the warped atlas intensity image after nonrigid registration.[13] Similar to the work for brain MR segmentation,[15] the NMI is computed within a ROI, defined by a mask image computed from the endo- and epicardial surfaces of the segmentation result, after a morphological dilation of 5 mm.
(3) *NMI (affine REG)*: This method computes the NMI between the target and warped atlas images after a global affine registration.[25] This method does not require the computation-intensive nonrigid registration,[13] but the surfaces of the atlases are generally more misaligned with the ground truth surfaces. Therefore, the ROI mask is computed using a morphological dilation operation of 10 mm, to cover the ground truth surfaces of the target image.
(4) *Random (no ranking)*: This method does not rank the atlases and only *randomly* selects a subset of atlases from the database for registration and label fusion.

3.A.3. Label fusion

Majority voting fusion (MVF) is a straightforward and widely used algorithm to fuse multiple classifiers.[26] This method is computationally the most efficient, but it does not account for the different performance of each atlas and the misalignment in atlas-to-target registration. Hence,

the performance of MAS using MVF can be limited. In this work, we employed JLF, which estimated local weights by minimizing the expectation of labeling errors and adopted a patch-based strategy to correct the local misalignment.[35] The patch size and search range are the important parameters in determining the performance of label fusion. We employed three patch sizes and four patch search ranges for comparisons in the experiment, including $3\times3\times3$, $5\times5\times5$, and $7\times7\times7$ mm for the patch size, and $1\times1\times1$, $3\times3\times3$, $5\times5\times5$, and $7\times7\times7$ mm for the search range.

To study the performance of MAS using different label fusion algorithms, we compared the segmentation accuracy of JLF with three label fusion methods, including (1) MVF, (2) locally weighted voting fusion (LWF), and (3) PPF.[21] LWF is a special case of PPF where the patch size is the size of a pixel and the search range is limited to the target pixel.[21] In our experiments, the patch size and search range for JLF were, respectively, set to $7\times7\times7$ and $5\times5\times5$ mm, and the parameters for PPF were both set to $3\times3\times3$ mm. These were the optimal parameters for WHS of CTA data based on our previous experience.

3.B. Segmentation evaluation

Dice score ($DS = 2|V \cap U|/(|V| + |U|)$),[26] Jaccard index ($JI = |V \cap U|/(|V \cup U|)$),[42] and surface-to-surface distance were used to assess the segmentation accuracy of a substructure. In this formulation, V and U indicate two segmentation results. Surface distance (SD) is computed from average distance between the automatically segmented surface and the manually delineated surface,

$$SD = \frac{\sum_{x_1\in\Omega_1} D(x_1,C_2) + \sum_{x_2\in\Omega_2} D(x_2,C_1)}{|\Omega_1| + |\Omega_2|}, \quad (4)$$

where C_1 and C_2 denote the surfaces of two segmentation results, Ω_1 and Ω_2 are the sets of sample points from C_1 and C_2, respectively; $D(x_1,C_2)$ is the Euclidean distance of x_1 to a triangle formed by three points in C_2 which are the closest to x_1, and $D(x_2,C_1)$ is computed in a similar manner.

To assess the volume accuracy of the WHS, the generalized overlap measures were employed,[3,41]

$$DC = \frac{2\sum_{s=1}^{7}|V_s \cap U_s|}{\sum_{s=1}^{7}(|V_s| + |U_s|)}, \quad JI = \frac{\sum_{s=1}^{7}|V_s \cap U_s|}{\sum_{s=1}^{7}|V_s \cup U_s|}, \quad (5)$$

where V_s and U_s denote the two segmentation results of a substructure. Similarly, we computed a generalized distance metric of the seven evaluated surfaces to indicate a WHS SD, as follows:

$$SD = \frac{\sum_{s=1}^{7}\left(\sum_{x_1\in\Omega_{s_1}} D(x_1,C_{s_2}) + \sum_{x_2\in\Omega_{s_2}} D(x_2,C_{s_1})\right)}{\sum_{s=1}^{7}(|\Omega_{s_1}| + |\Omega_{s_2}|)}. \quad (6)$$

Here, the distance, $D(x_1,C_{s_2})$ and $D(x_2,C_{s_1})$, is computed between the surfaces of corresponding substructures of the two WHS results. Instead of computing SD directly from the two whole heart surfaces using Eq. (4), the metric using Eq. (6) provides a more objective measurement for the evaluation of WHS.

3.C. Experiment design

Thirty clinical cardiac CTA images, described in Sec. 2, were used. The proposed MAS adopted the hierarchical scheme for the target-to-image registration, the proposed conditional entropy for atlas ranking, and the JLF for label fusion. A leave-one-out cross validation strategy was adopted, where each subject was considered as a test case and the remaining 29 were used to form an atlas database. The 11 best ranked atlases were selected for label fusion.

For comparisons, results of the SAS were reported. For SAS, each of the 30 subjects was treated as an atlas to segment the other 29 images, resulting in $30\times29 = 870$ SAS cases.

Three comparison studies were performed:

(1) MAS using different label fusion strategies, including different label fusion methods and different parameterizations of the patch-based strategy for JLF;
(2) MAS using different atlas ranking methods;
(3) MAS using different sizes of atlas databases.

The algorithms used in this work were implemented on a Lenovo ThinkStation D30 workstation. The implementations were based on single thread for the atlas-to-target registration in a SAS. In MAS experiments, the atlas-to-target registration tasks were run simultaneously using four Intel Xeon E5-2667V2 CPUs, which had 32 cores in total.

3.C.1. Study of atlas ranking and selection

In this study, MAS using the proposed conditional entropy and nonrigid registration was compared with the four schemes described in Sec. 3.A.2. All the five MAS schemes adopted MVF for label fusion. A leave-one-out cross validation strategy was used. For each MAS scheme, we analyzed the performance curves of the mean generalized Dice scores of WHS with respect to the number of atlases selected for segmentation and label fusion. To objectively assess the effectiveness of an atlas ranking method, we investigated the performance curves according to three criteria:

(1) the Dice score of the MAS scheme using the *one* best ranked atlas;
(2) the number of atlases needed to reach the optimal performance;
(3) the Dice score of the optimal MAS performance.

3.C.2. Study of influence of atlas database

We investigated the performance of a MAS scheme with respect to the size of atlas databases in two aspects: First, we analyzed the influence of the size of an atlas database on the performance of a MAS scheme; second, we identified the number of atlases required for a MAS scheme to reach the optimal performance.

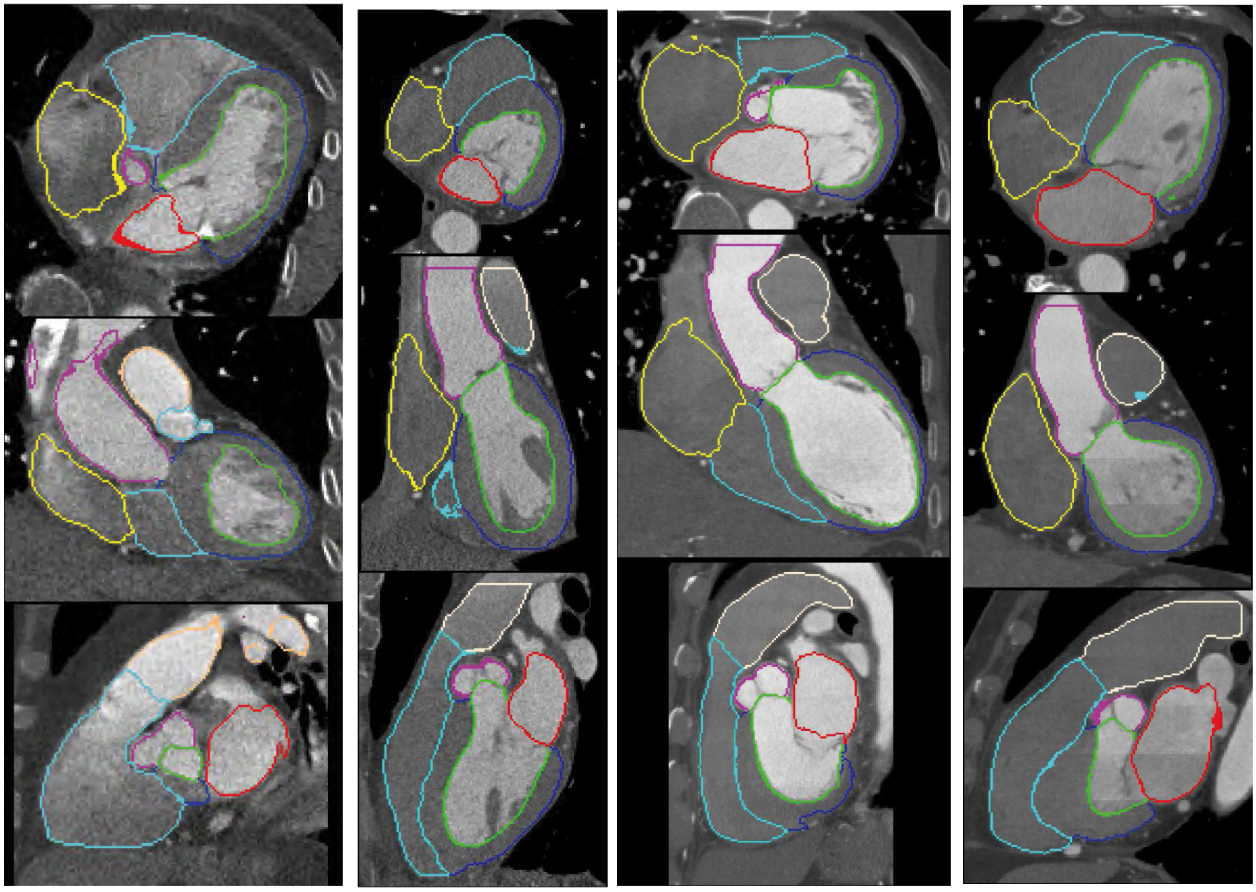


FIG. 4. Three orthogonal views of four segmentation cases with the WHS Dice scores in brackets, illustrating performance of the proposed MAS scheme: case 1 and case 2 are the two *worst* cases among all the test subjects in terms of WHS Dice scores, and case 3 and case 4 are the two median cases.

We randomly selected a certain number of atlases from the available 30 subjects to form an atlas database. The remaining subject(s) were then used to form a test data set. In our study, we chose one third (referred to as *Size 11*), half (referred to as *Size 15*), two thirds (referred to as *Size 20*), three quarters (referred to as *Size 25*), and all but one (referred to as *Size 29*) from the 30 subjects, resulting in five different sizes. For *Size 11*, we employed 114 cases for the experiments; for *Size 15*, we used 90 cases; for both *Size 20* and *Size 25*, we selected 60 MAS cases; and for *Size 29*, we employed the leave-one-out strategy.

The MAS scheme adopted the proposed atlas ranking algorithm and the MVF method for label fusion.

4. RESULTS

4.A. Performance of the proposed MAS scheme

Figure 4 provides visual results of four cases: the three orthogonal views (axial, coronary, and sagittal views) of the CT images superimposed with the contours of automatic segmentation using the proposed WHS scheme. Case-1 and case-2 are the two worst segmentation results among the 30 test subjects in terms of WHS Dice score, and case-3 and case-4 are the two median cases.

Table II provides the segmentation accuracies of the SAS and the proposed MAS scheme, including the Dice scores,

TABLE II. Performance of the SAS and the proposed MAS scheme; SD: surface distance; Myo: myocardium; Epi: epicardium.

Dice	LV	Myo	RV	LA	RA	Aorta	PA	WHS
SAS	0.910 ± 0.0546	0.760 ± 0.0955	0.829 ± 0.0879	0.842 ± 0.0817	0.761 ± 0.121	0.867 ± 0.0989	0.699 ± 0.165	0.825 ± 0.0670
MAS	0.967 ± 0.0125	0.901 ± 0.0242	0.918 ± 0.0308	0.920 ± 0.0413	0.872 ± 0.0481	0.955 ± 0.0153	0.789 ± 0.1294	0.918 ± 0.0212
Jaccard	LV	Myo	RV	LA	RA	Aorta	PA	WHS
SAS	0.838 ± 0.0830	0.621 ± 0.113	0.716 ± 0.113	0.735 ± 0.113	0.627 ± 0.135	0.776 ± 0.129	0.559 ± 0.180	0.708 ± 0.0880
MAS	0.936 ± 0.0227	0.821 ± 0.0391	0.849 ± 0.0511	0.855 ± 0.0685	0.775 ± 0.0739	0.915 ± 0.0278	0.669 ± 0.1685	0.849 ± 0.0357
SD (mm)	LV	Epi	RV	LA	RA	Aorta	PA	WHS
SAS	1.35 ± 1.20	1.90 ± 1.85	2.20 ± 2.20	2.02 ± 2.08	2.98 ± 2.87	1.19 ± 1.29	1.68 ± 1.58	1.97 ± 2.27
MAS	0.82 ± 0.55	1.10 ± 0.76	1.56 ± 1.03	1.79 ± 1.08	2.46 ± 1.64	0.84 ± 0.47	1.35 ± 0.79	1.58 ± 0.92

TABLE III. The WHS Dice scores and runtime (minute) of the MAS schemes using different label fusion algorithms.

	MAS-MVF	MAS-LWF	MAS-PPF	MAS-JLF
Dice	0.901 ± 0.0276	0.905 ± 0.0247	0.909 ± 0.0249	0.918 ± 0.0212
Time	2.82	2.89	3.62	13.15

Jaccard indices, and SD of each substructure as well as the whole heart. The accuracies of the proposed MAS were better than those of the SAS for all categories with statistical significance using the paired, two-tailed t-test ($p < 0.001$). Note that in the MAS, the computation of the conditional entropy ranking criterion was based on the whole heart. Hence, the segmentation accuracy for a specific substructure could be further improved, if the conditional entropy was computed solely based on the substructure(s) of interest. For the runtime, the SAS took 2.82 min on average for a case and the MAS took 13.15 min.

4.B. Study of label fusion strategies

Table III provides the WHS Dice scores and runtime of the four MAS schemes described in Sec. 3.A.3, i.e., MAS-MVF, MAS-LWF, MAS-PPF, and MAS-JLF. To assess the significance of performance difference, we used the two-tail, paired student t-test. MAS-JLF was statistically significantly better than MAS-PPF ($p < 0.0001$), MAS-PPF was significantly better than MAS-LWF ($p < 0.0001$), and MAS-LWF was significantly better than MAS-MVF ($p = 0.000\,86$). Using the Bonferroni correction, we obtained the performance ranking of the four atlas ranking methods, from the worst to the best as MVF, LWF, PPF, JLF, with statistical significance ($p < 0.001$).

Figure 5 plots the WHS Dice scores of the MAS-JLF using different patch-based strategies, as described in Sec. 3.A.3. For the patch search range, JLF did not demonstrate evident difference by increasing the search range after $3\times3\times3$ mm. Also, when the patch size was small ($3\times3\times3$ mm), the fusion results could be indeed poor, compared to the results using patch size $5\times5\times5$ or $7\times7\times7$ mm. The optimal parameters were $7\times7\times7$ mm for the patch size and $5\times5\times5$ mm for the search range.

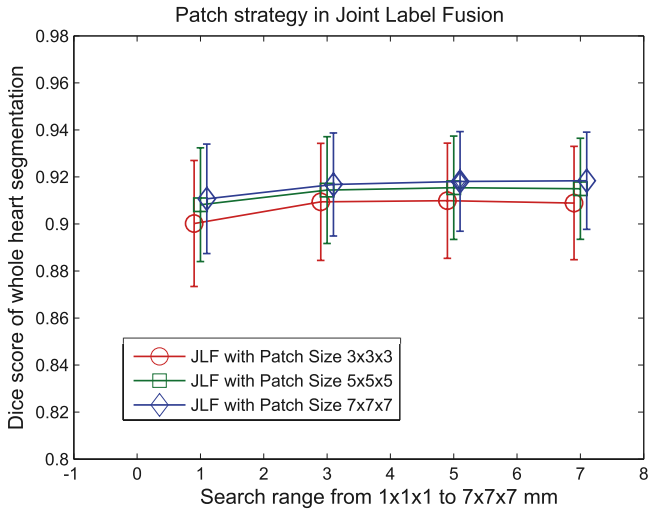


FIG. 5. The WHS Dice scores of the MAS using joint label fusion (JLF) with different patch sizes, from $3\times3\times3$ to $7\times7\times7$ mm, and patch-search ranges, from $1\times1\times1$ to $7\times7\times7$ mm.

4.C. Study of atlas ranking and selection

Figure 6(a) plots the Dice score curves of the five MAS schemes described in Sec. 3.A.2. The mean WHS Dice score of SAS, 0.825 ± 0.0670, is also presented. The curve of MAS *Conditional entropy (nonrigid REG)* was the closest to that of MAS *Manual ranking (ground truth)*, indicating that the proposed atlas ranking scheme produced the best performance compared to the other four methods.

In addition, we assessed the effectiveness of the atlas ranking using the three criteria described in Sec. 3.C. First, when selecting one atlas, four of the five MAS schemes, except for MAS *Random (no ranking)*, produced better Dice scores than the SAS, indicating that these four atlas ranking methods were effective and improved the MAS segmentation. The performance rankings of them were as follows: MAS *Manual ranking (ground truth)*, MAS *Conditional entropy (nonrigid REG)*, MAS *NMI (nonrigid REG)*, and MAS *NMI (affine REG)*.

Second, all the five curves demonstrated a global optimum. In general, the performance curve of a MAS scheme using an effective atlas ranking approach should improve rapidly. The fewer the number of atlases for the curve to reach the optimum, the better the atlas ranking method is considered. As Fig. 6(a) shows, the performance rankings were as follows: MAS *Manual ranking (ground truth)*, MAS *Conditional entropy (nonrigid REG)*, MAS *NMI (nonrigid REG)*, MAS *NMI (affine REG)*, and MAS *Random (no ranking)*.

Finally, the performance rankings according to the optimal Dice scores of the five schemes were as follows: MAS *Manual ranking (ground truth)*, MAS *Conditional entropy (nonrigid REG)*, MAS *NMI (nonrigid REG)*, MAS *NMI (affine REG)*, and MAS *Random (no ranking)*. The WHS Dice scores of the five MAS schemes using 11 atlases, where all the performance curves had started converging, were, respectively, 0.905 ± 0.025, 0.901 ± 0.028, 0.898 ± 0.030, 0.896 ± 0.030, and 0.890 ± 0.034. MAS *Conditional entropy (nonrigid REG)* was significantly better than MAS *NMI (nonrigid REG)* ($p = 0.028$), *NMI (affine REG)* ($p < 0.001$), and MAS *Random (no ranking)* ($p < 0.001$). This result agrees with the two conclusions drawn above and confirms the effectiveness of the atlas ranking methods.

4.D. Study of influence of atlas database

Figure 6(b) shows the performance of MAS using different sizes of atlas databases. Each curve shows a global optimum with respect to the number of atlases selected for label fusion.

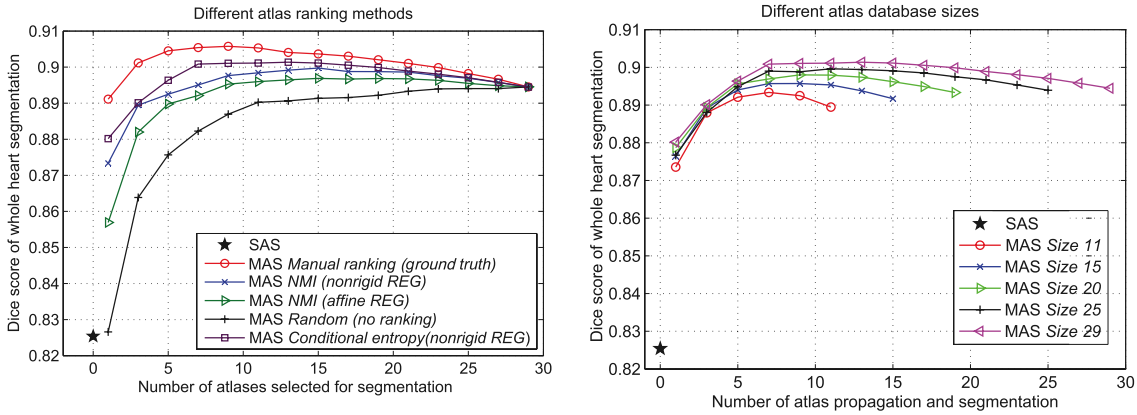


FIG. 6. Performance of the MAS schemes using different atlas ranking methods (a) and different atlas databases (b); the result of SAS is provided for comparisons.

In addition, the numbers of atlases for the curves to reach their optima are different. In general, the smaller the atlas database is, the fewer number of atlases is needed to reach the optimum performance. For example, the numbers for the MAS to reach the global optimal performance were, respectively, 7, 9, 13 for atlas databases of size 11, 20, 29. Finally, MAS using a larger atlas database generated more accurate WHS results than the MAS using a smaller database when the same number of atlases was used for label fusion.

5. DISCUSSION

In the label fusion study, LWF improved the WHS Dice score to 0.905 ± 0.025, compared to 0.901 ± 0.028 in the case of using traditional MVF, and the result was comparable to the accuracy of MAS-MVF using *Manual ranking (ground truth)* in Fig. 6(a). PPF further improved the segmentation accuracy to 0.909 ± 0.025 by using the patch-based strategy, and JLF achieved the best result, 0.918 ± 0.021, by tackling correlated labeling errors in different atlases. The improvements were statistically significant ($p < 0.001$) using the Bonferroni correction. In the patch search strategy, the label fusion performance was not necessarily improved by increasing the patch size and search range. This is probably due to the fact that the intensities between different substructures, such as LV and LA, can be similar and thus cause labeling errors in patch-based decision making.

In the atlas ranking experiment, the proposed conditional entropy method yielded the highest Dice score compared to the conventional intensity-based method. This is because our method provides a mechanism to incorporate atlas segmentation into the evaluation of atlas ranking and thus avoids biased rankings due to the inhomogeneous intensity patterns which are commonly presented in cardiac CTA images.

Section 4.D shows that MAS benefits more from a larger atlas database. However, when the number of atlases is significantly greater than the number of cores of the workstation, it is recommended to adopt a *multilevel* atlas ranking scheme, by first using a fast atlas ranking method to select a subset of *good* atlases in order to build a subject-specific atlas database. Then, a MAS scheme with a more effective atlas ranking algorithm can be applied using the smaller but more *subject-specific* atlas database. For example, the *NMI (affine REG)* method only takes 4 s for a single thread process to rank an atlas, 3% runtime compared to that of *Conditional entropy (nonrigid REG)*. Therefore, one can select 30 subject-specific atlases from a *much larger* atlas pool using *NMI (affine REG)* ranking method, and then applies *Conditional entropy (nonrigid REG)* to achieve the optimal MAS performance without sacrificing computational efficiency. In future work, we will investigate

TABLE IV. The segmentation results reported in the literature. N/A indicates the segmentation error was not provided; Myo/Epi: myocardium for Dice or epicardium for SD.

	Zheng *et al.* (Ref. 9)	Ecabert *et al.* (Ref. 43)	Kirisli *et al.* (Ref. 11)	
Error	SD (mm)	SD (mm)	SD (mm)	Dice
LV	1.13 ± 0.55	0.77 ± 1.14	0.62 ± 0.63	0.95
Myo/Epi	1.21 ± 0.42	0.68 ± 0.96	1.04 ± 1.15	N/A
RV	1.55 ± 0.38	0.63 ± 0.66	1.40 ± 1.47	0.90
LA	1.32 ± 0.42	0.70 ± 0.87	0.66 ± 0.84	0.94
RA	1.57 ± 0.48	0.82 ± 1.00	1.44 ± 1.88	0.89
Aorta	N/A	0.60 ± 1.14	0.44 ± 0.60	0.94
PA	N/A	0.50 ± 0.49	N/A	N/A
WHS	N/A	0.82 ± 1.00	0.99 ± 1.25	N/A
Runtime	4 s	10–30 s	20 min	

this multilevel atlas ranking scheme using a large database constructed from multiple centers.

Table IV presents the WHS results of cardiac CT in the literature. Zheng *et al.*[9] validated their algorithm using more than 500 CT images from 137 subjects, among which 242 were used for training and 323 were used for testing. The training data and test data could come from the same subjects. Ecabert *et al.*[43] employed a leave-one-out cross validation strategy on a cohort of 28 subjects. They reported the surface distance error based on a triangulated surface mesh model, which was used for the construction of both the gold standard segmentation and the automatic segmentation. Kirisli *et al.*[11] performed a multicenter, multivendor evaluation of a conventional MAS method. The whole heart validation measurements reported were obtained from the leave-one-out cross validation of eight subjects. From the reported results, Ecabert *et al.*[43] achieved the best surface distance accuracy, while the method developed by Zheng *et al.*[9] was the most efficient in terms of runtime. However, note that due to the difference in the test data sets, evaluation metrics and computer hardwares, an objective comparison between these works can be difficult.

6. CONCLUSIONS

This work presents a new MAS method for the WHS of CT data. The MAS adopts a hierarchical registration scheme specifically designed for cardiac images, a new atlas ranking method based on conditional entropy, and JLF for label fusion. We evaluated this WHS method and investigated different segmentation strategies using 30 clinical cardiac CTA images. The proposed MAS method yielded a mean WHS Dice score of 0.918 ± 0.21, a Jaccard index of 0.849 ± 0.036, a surface distance of 1.58 ± 0.92 mm, and the mean computation time for one case was 13.2 min. The new atlas ranking criterion was shown to outperform the conventional intensity-based schemes, and the MAS using a larger atlas database achieved better segmentation accuracy (WHS Dice scores) than using a smaller one. Our WHS method is fully automatic and is robust and accurate for geometrical modeling of the heart. It can be useful in clinic care of cardiovascular diseases.

ACKNOWLEDGMENTS

This work was supported by the Chinese NSFC research fund (No. 81301283) and the SRF for ROCS, SEM of China (47th).

[a]Author to whom correspondence should be addressed. Electronic mail: zhuangxiahai@sjtu.edu.cn

[1]World Health Organization, "Cardiovascular diseases (CVDs) Fact Sheet No. 317," http://www.who.int/mediacentre/factsheets/fs317/en/index.html (2011).
[2]C. Chartrand-Lefebvre, A. Cadrin-Chênevert, E. Bordeleau, P. Ugolini, R. Ouellet, J.-L. Sablayrolles, and J. Prenovault, "Coronary computed tomography angiography: Overview of technical aspects, current concepts, and perspectives," Can. Assoc. Radiol., J. **58**, 92–108 (2007).
[3]X. Zhuang, "Challenges and methodologies of fully automatic whole heart segmentation: A review," J. Healthcare Eng. **4**, 371–407 (2013).
[4]C. Petitjean and J. N. Dacher, "A review of segmentation methods in short axis cardiac MR images," Med. Image Anal. **15**, 169–184 (2011).
[5]D. Kang, J. Woo, P. J. Slomka, D. Dey, G. Germano, and C.-C. Jay Kuo, "Heart chambers and whole heart segmentation techniques: Review," J. Electron. Imaging **21**, 010901 (2012).
[6]R. S. Vasan, E. J. Benjamin, and D. Levy, "Congestive heart failure with normal left ventricular systolic function clinical approaches to the diagnosis and treatment of diastolic heart failure," Arch. Intern. Med. **156**, 146–157 (1996).
[7]J. Lotjonen, S. Kivisto, J. Koikkalainen, D. Smutek, and K. Lauerma, "Statistical shape model of atria, ventricles and epicardium from short- and long-axis mr images," Med. Image Anal. **8**, 371–386 (2004).
[8]J. Koikkalainen, T. Tolli, K. Lauerma, K. Antila, E. Mattila, M. Lilja, and J. Lotjonen, "Methods of artificial enlargement of the training set for statistical shape models," IEEE Trans. Med. Imaging **27**, 1643–1654 (2008).
[9]Y. Zheng, A. Barbu, B. Georgescu, M. Scheuering, and D. Comaniciu, "Four-chamber heart modeling and automatic segmentation for 3-D cardiac CT volumes using marginal space learning and steerable features," IEEE Trans. Med. Imaging **27**, 1668–1681 (2008).
[10]J. Peters, O. Ecabert, C. Meyer, R. Kneser, and J. Weese, "Optimizing boundary detection via simulated search with applications to multi-modal heart segmentation," Med. Image Anal. **14**, 70–84 (2009).
[11]H. Kirisli, M. Schaap, S. Klein, S. L. Papadopoulou, M. Bonardi, C. H. Chen, A. C. Weustink, N. R. Mollet, E. J. Vonken, R. J. van der Geest, T. van Walsum, and W. J. Niessen, "Evaluation of a multi-atlas based method for segmentation of cardiac CTA data: A large-scale, multicenter, and multivendor study," Med. Phys. **37**, 6279–6291 (2010).
[12]X. Zhuang, K. Rhode, R. Razavi, D. J. Hawkes, and S. Ourselin, "A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI," IEEE Trans. Med. Imaging **29**, 1612–1625 (2010).
[13]T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer, "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains," NeuroImage **21**, 1428–1442 (2004).
[14]T. Rohlfing and J. C. R. Maurer, "Multi-classifier framework for atlas-based image segmentation," Pattern Recognit. Lett. **26**, 2070–2079 (2005).
[15]P. Aljabar, R. Heckemann, A. Hammers, J. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy," NeuroImage **46**, 726–738 (2009).
[16]R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," NeuroImage **33**, 115–126 (2006).
[17]E. van Rikxoort, I. Isgum, Y. Arzhaeva, M. Staring, S. Klein, M. Viergever, J. Pluim, and B. van Ginneken, "Adaptive local multi-atlas segmentation application to the heart and the caudate nucleus," Med. Image Anal. **14**, 39–49 (2010).
[18]T. Langerak, U. van der Heide, A. Kotte, M. Viergever, M. van Vulpen, and J. Pluim, "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple)," IEEE Trans. Med. Imaging **29**, 2000–2008 (2010).
[19]B. Ballanger, L. Tremblay, V. Sgambato-Faure, M. Beaudoin-Gobert, F. Lavenne, D. Le Bars, and N. Costes, "A multi-atlas based method for automated anatomical Macaca fascicularis brain MRI segmentation and PET kinetic extraction," NeuroImage **77**, 26–43 (2013).
[20]A. Gholipour, A. Akhondi-Asl, J. A. Estroff, and S. K. Warfield, "Multi-atlas multi-shape segmentation of fetal brain MRI for volumetric and morphometric analysis of ventriculomegaly," NeuroImage **60**, 1819–1831 (2012).
[21]W. Bai, W. Shi, D. P. O'Regan, T. Tong, H. Wang, S. Jamil-Copley, N. S. Peters, and D. Rueckert, "A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: Application to cardiac MR images," IEEE Trans. Med. Imaging **32**, 1302–1315 (2013).
[22]S. Zhang, Y. Zhan, and D. N. Metaxas, "Deformable segmentation via sparse representation and dictionary learning," Med. Image Anal. **16**, 1385–1396 (2012).
[23]W. M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximisation of mutual information," Med. Image Anal. **1**, 35–51 (1996).
[24]F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," IEEE Trans. Med. Imaging **16**, 187–198 (1997).
[25]C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment," Pattern Recognit. **32**, 71–86 (1999).
[26]J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," IEEE Trans. Pattern Anal. Mach. Intell. **20**, 226–239 (1998).

[27]X. Artaechevarria, A. Munoz-Barrutia, and C. O. de Solorzano, "Combination strategies in multi-atlas image segmentation: Application to brain mr data," IEEE Trans. Med. Imaging **28**, 1266–1277 (2009).
[28]I. Isgum, M. Staring, A. Rutten, M. Prokop, M. Viergever, and B. van Ginneken, "Multi-atlas-based segmentation with local decision fusion–Application to cardiac and aortic segmentation in ct scans," IEEE Trans. Med. Imaging **28**, 1000–1010 (2009).
[29]P. Coupe, J. V. Manjon, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins, "Nonlocal patch-based label fusion for hippocampus segmentation," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2010*, Lecture Notes in Computer Science Vol. 6363 (Springer, Berlin, Heidelberg, 2010), pp. 129–136.
[30]M. Sabuncu, B. Yeo, K. V. Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion," IEEE Trans. Med. Imaging **29**, 1714–1729 (2010).
[31]R. Wolz, C. Chu, K. Misawa, K. Mori, and D. Rueckert, "Multi-organ abdominal ct segmentation using hierarchically weighted subject-specific atlases," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2012*, Lecture Notes in Computer Science Vol. 7510 (Springer, Berlin, Heidelberg, 2012), pp. 10–17.
[32]S. F. Eskildsen, P. Coupe, V. Fonov, J. V. Manjon, K. K. Leung, N. Guizard, S. N. Wassef, L. R. Qstergaard, and D. L. Collins, "BEaST: Brain extraction based on nonlocal segmentation technique," NeuroImage **59**, 2362–2373 (2012).
[33]R. Wolz, C. Chu, K. Misawa, K. Mori, and D. Rueckert, "Multi-organ abdominal CT segmentation using hierarchically weighted subject-specific atlases," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2012*, Lecture Notes in Computer Science Vol. 15 (Springer, Berlin, Heidelberg, 2012), pp. 10–17.
[34]J. E. Iglesias, M. R. Sabuncu, and K. Van Leemput, "A unified framework for cross-modality multi-atlas segmentation of brain MRI," Med. Image Anal. **17**, 1181–1191 (2013).
[35]H. Wang, J. W. Suh, S. R. Das, J. Pluta, C. Craige, and P. A. Yushkevich, "Multi-atlas segmentation with joint label fusion," IEEE Trans. Pattern Anal. Mach. Intell. 611–623 (2013).
[36]S. Hu, P. Coupe, J. C. Pruessner, and D. L. Collins, "Nonlocal regularization for active appearance model: Application to medial temporal lobe segmentation," Hum. Brain Mapp. **35**, 377–395 (2014).
[37]P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," NeuroImage **31**, 1116–1128 (2006).
[38]D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," IEEE Trans. Med. Imaging **18**, 712–721 (1999).
[39]X. Zhuang, K. Rhode, S. Arridge, R. Razavi, D. Hill, D. Hawkes, and S. Ourselin, "An atlas-based segmentation propagation framework using locally affine registration–Application to automatic whole heart segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, Lecture Notes in Computer Science Vol. 5242 (Springer, Berlin, Heidelberg, 2008), pp. 425–433.
[40]G. Sanroma, G. Wu, Y. Gao, and S. Dinggang, "Learning to rank atlases for multiple-atlas segmentation," IEEE Trans. Med. Imaging **33**(10), 1939–1953 (2014).
[41]W. R. Crum, O. Camara, and D. L. G. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," IEEE Trans. Med. Imaging **25**, 1451–1461 (2006).
[42]P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," Bulletin de la Société Vaudoise des Sciences Naturelles **37**, 547–579 (1901).
[43]O. Ecabert, J. Peters, H. Schramm, C. Lorenz, J. von Berg, M. J. Walker, M. Vembar, M. E. Olszewski, K. Subramanyan, G. Lavi, and J. Weese, "Automatic model-based segmentation of the heart in CT images," IEEE Trans. Med. Imaging **27**, 1189–1201 (2008).