A Probabilistic Patch-Based Label Fusion Model for Multi-Atlas Segmentation with Registration Refinement: Application to Cardiac MR Images

Wenjia Bai, Wenzhe Shi, Declan P. O'Regan, Tong Tong, Haiyan Wang, Shahnaz Jamil-Copley, Nicholas S. Peters* and Daniel Rueckert

***Abstract*—The evaluation of ventricular function is important for the diagnosis of cardiovascular diseases. It typically involves measurement of the left ventricular (LV) mass and LV cavity volume. Manual delineation of the myocardial contours is time-consuming and dependent on the subjective experience of the expert observer. In this paper, a multi-atlas method is proposed for cardiac MR image segmentation. The proposed method is novel in two aspects. First, it formulates a patch-based label fusion model in a Bayesian framework. Second, it improves image registration accuracy by utilising label information, which leads to improvement of segmentation accuracy. The proposed method was evaluated on a cardiac MR image set of 28 subjects. The average Dice overlap metric of our segmentation is 0.92 for the LV cavity, 0.89 for the RV cavity and 0.82 for the myocardium. The results show that the proposed method is able to provide accurate information for clinical diagnosis.**

***Index Terms*—image segmentation, multi-atlas segmentation, patch-based segmentation, image registration**

I. Introduction

The evaluation of ventricular function is important for the diagnosis of cardiovascular diseases, such as hypertrophic cardiomyopathy (HCM), ischaemic heart disease (IHD), arrhythmogenic right ventricular displasia (ARVD) etc. The evaluation normally involves the measurement of ventricular mass (VM), end-diastolic volume (EDV) and end-systolic volume (ESV) from cardiac MR images. Conventionally, quantification of these parameters mainly relies on manual delineation of the myocardial contours, which is time-consuming and dependent on the observer's experience. Therefore, a great number of semi-automatic or automatic approaches have been developed for cardiac image segmentation [1]–[8]. Left ventricle (LV) segmentation challenge competitions have been held twice, respectively in 2009 and 2011 during the MICCAI conferences [9], [10]. In addition, last year (2012), the MICCAI conference has hosted a segmentation challenge competition for the right ventricle (RV) [11], [12].

W. Bai, W. Shi, T. Tong, H. Wang and D. Rueckert are with Biomedical Image Analysis Group, Dept. of Computing, Imperial College London, UK.

D.P. O'Regan is with MRC Clinical Sciences Centre, Robert Steiner MRI Unit, Hammersmith Hospital, Imperial College London, UK.

S. Jamil-Copley and N.S. Peters are with Cardiology Department, St Marys Hospital, Imperial College Healthcare NHS Trust, UK.

*Corresponding author: N.S. Peters (n.peters@imperial.ac.uk).

Multi-atlas segmentation has been demonstrated to significantly improve segmentation accuracy compared to segmentation based on a single atlas [13]–[16]. It utilises a set of atlases that represent the inter-subject variability of the cardiac anatomy. The target image to be segmented is registered to each atlas and the propagated labels from multiple atlases are combined to form a consensus segmentation. This method has two advantages compared to single atlas propagation. First, it accounts for the anatomical shape variability by using multiple atlases. Second, it is robust because segmentation errors associated with single atlas propagation can be averaged out when combining multiple atlases. The consensus segmentation is less likely to be affected, when an individual atlas does not match the target image well or when serious registration errors occur for an individual atlas. Its performance can be improved by selecting a subset of atlases which look more similar to the target image than the other images [17]–[19], or by using weighted label fusion in which the contribution of each atlas is either globally or locally weighted by its similarity to the target image [20]–[23] or an estimated performance level of this atlas [24].

Conventionally, the label at a voxel in the target image is determined by combining the labels at the corresponding voxels from each atlas. It assumes that the target image and each atlas image has been accurately registered. Recently, Coupé et al. proposed the patch-based segmentation method[1] [25], [26]. The proposed method do not require non-rigid image registration and only affine registration is performed to align the atlas with the target. To account for image registration errors, a number of patches (image sub-volumes) near the atlas voxel are considered for label fusion. The underlying assumption is that if the atlas patch and the target patch look similar, they should also have a similar label. The similarity between the target patch and the atlas patch is used to determine the weights for the label fusion process. Patch-based segmentation borrows the idea from non-local[2] means image denoising [27]–[29], which assumes that image information is redundant. Therefore, an image patch can have similar patches

[1]In this paper, we consider the patch-based segmentation method as a subcategory of the multi-atlas segmentation method, in which multiple patches are used to weight the label fusion process.

[2]"Non-local" denotes that the patch weight and the estimation support is only dependent on intensity similarity and is regardless of its distance from the patch to the voxel of interest. It follows the nomenclature widely established in the literature on image denoising and restoration [27].

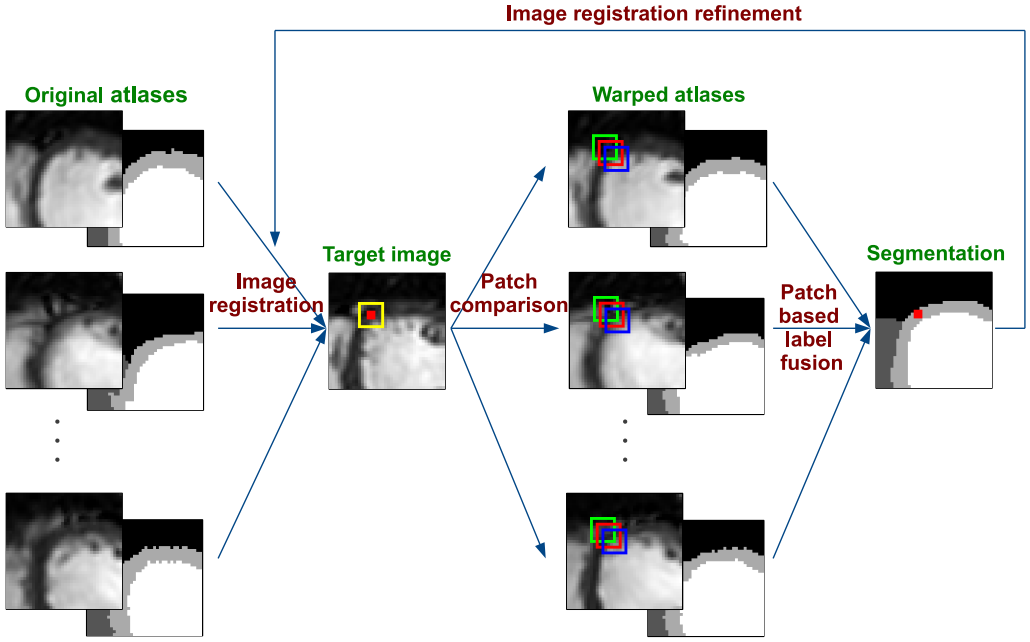


Fig. 1: The atlases and corresponding label maps are warped to the target image space using image registration. In order to determine the label at a voxel (the red dot), a patch (the yellow rectangle) in the target image is compared to a number of patches (the colourful rectangles) in the warped atlases. Each atlas patch is weighted according to its similarity and distance to the target patch. The labels from all the atlas patches are combined to give the label estimate in the target image. The resulting segmentation can be incorporated back into the image registration process to refine the registration results.

in the same image, which can even be very distant from the first patch. The intensities of these patches are then averaged for denoising. In the context of multi-atlas segmentation, since all the atlases are at least roughly registered to the target image, we only need to search for similar patches in a small neighbourhood, instead of in the whole image.

Since then, numerous efforts have been made to improve the performance of the patch-based segmentation method and to apply it to various images. Rousseau et al. [30] proposed a multi-point model for label fusion and compared it to the point-wise model. Wang et al. [31] estimated the patch weight by polynomial regression, where a regression model relates the label to a function of the patch intensities. Tong et al. [32] and Zhang et al. [33] respectively proposed to pursue a sparse representation of the target patch from a dictionary of atlas patches by minimising the L_1-norm. Wang et al. [34] proposed to correct for systematic errors by building two classifiers, respectively for error detection and error correction. This method yielded the best published results for hippocampus segmentation so far. Recently, Wang et al. [35] also proposed to estimate the optimal weight for label fusion by minimising the expectation of the labelling error. Zhang et al. [36] proposed to use an adaptive Markov random field (MRF) model based on a map of registration confidence. Hu et al. [37] applied patch-based segmentation to refine an initial segmentation produced by an active appearance model (AAM). Wolz et al. [38] proposed a hierarchical framework for label fusion and applied it to multi-organ abdominal CT segmentation. Fonov et al. [39] applied the patch-based method to segmentation of brain structures. Eskildsen et al. [40] used the patch-based method for brain extraction with a multi-resolution framework to accelerate computation.

Despite the abundant literature on patch-based segmentation, most of the efforts have been focused on improving the strategy for patch weight estimation [31]–[33], [35] or customising the patch-based method to specific applications [37]–[40]. However, a statistical model for patch-based label fusion is still missing. Warfield et al. [24] pioneered early work in this direction by proposing a Bayesian model for integration of multiple segmentations. In this approach, each segmentation is essentially regarded as a rater and the performance (sensitivity and specificity) of each rater is estimated in the EM framework. This method assigns a global weight to each rater during label fusion and atlas intensity information is not accounted by the model. Therefore, it is not applicable to patch-based label fusion in its original form. Recently, Sabuncu et al. [23] proposed a generative model for multi-atlas label fusion. Iglesias et al. [41], [42] refined this model to account for multi-modal images by incorporating a Gaussian mixture model. This model accounts for intensity information of each atlas. However, label fusion is performed accounting for each single voxel only, without considering any patches. Again, this is not a patch-based model.

The first contribution of this paper is the extension of Sabuncu's work: we propose a probabilistic model for patch-based label fusion, which formulates the method in a statistical Bayesian framework. We have noticed that Asman et al. almost simultaneously proposed a statistical model for patch-based label fusion, the non-local STAPLE [43], [44]. The difference between Asman's model and our model is that Asman's model is based on STAPLE and replaces the global rater performance in STAPLE with local rater performance, which is estimated by patch-based similarity. In contrast, our model constructs the relationship between target patch and atlas patches by

introducing an auxiliary mapping field and analytically derive a solution to the maximisation of a posteriori probability (MAP) problem. In addition, our model considers not only patch-based similarity, but also accounts for the registration uncertainty between target and atlas, which also has an impact on label fusion.

The second contribution of this paper is that we improve the registration accuracy for each atlas by incorporating intermediate label information into image registration. We refer to this strategy as "registration refinement". The underlying rationale is that we regard multi-atlas label fusion as a classifier ensemble, in which each atlas acts as a classifier and the opinions from all the atlases are fused. Intuitively, if the registration performance of each atlas can be improved, the classifier ensemble should also perform better. Some groups have proposed to use non-rigid registration instead of affine registration in patch-based label fusion [30], [35], [38], [41], [42]. In this work, we incorporate label information into a non-rigid registration framework to improve accuracy, since discrete labels can be less ambiguous than intensity values in differentiating tissue types. Experiments have demonstrated that by introducing label information, the method outperforms image registration using intensity information only.

Our method was evaluated on a cardiac MR data set of 28 subjects. Experiments have shown that the method significantly improves the segmentation accuracy for the LV, RV and myocardium. Compared to ground truth segmentation, the Dice metrics of our method is 0.92 for the LV cavity, 0.89 for the RV cavity and 0.82 for the myocardium. We also measure the clinical indices commonly used for assessing ventricular function based on the automated segmentation. The results demonstrate good agreement with manual measurements. In addition, we compared our method to some other published methods and evaluated its performance on the MICCAI RV Segmentation Challenge data set [11], [12].

II. Methods

A. Framework

Consider an image $I = \{I(x)|x \in \Omega\}$, where x denotes the voxel and $\Omega \subset \mathbb{R}^3$ denotes the lattice on which the image is defined. The goal of segmentation is to estimate a label map L associated with the image I, in which each voxel is assigned a discrete label. The label takes discrete values from 1 to $\mathcal{L}$, representing a number of tissue classes.

In a multi-atlas segmentation method, we have a number of atlases $\{I_n|n = 1, \cdots, N\}$ with corresponding label maps $\{L_n|n = 1, \cdots, N\}$ already known. We register the target image I with each of the atlases I_n and infer a label map L'_n from this atlas. Combining the warped label maps from all the atlases, a fused label map is generated as the segmentation. Figure 1 illustrates the proposed method.

B. Patch-based Label Fusion Model

In this section, we present the probabilistic patch-based label fusion model. It is an extension of the generative model proposed by Sabuncu et al [23]. Sabuncu's model assumes that each voxel in the target image is generated from a corresponding voxel in one of the atlases. By introducing this one-to-one mapping from target voxel to atlas voxel, the multi-atlas segmentation problem can be nicely formulated in a Bayesian framework, where a posteriori probability is maximised. This model inherently assumes that the image registration is accurate, so that the mapping from the target voxel to the voxel in the warped atlas is valid.

However, in reality, image registration may not be perfect. If there is slight spatial mismatch between the target image and the warped atlas image, then the target voxel may correspond to a shifted position in the atlas. To account for this registration error, we consider a number of voxels in a local neighbourhood in the atlas, as the potential matches for the target voxel. In addition, we replace the voxels by patches since the computation of intensity similarity based on a patch may be more robust than that based on a single voxel. The label in the target image is determined by comparing the target patch to atlas patches and then combining the patch labels. This is a patch-based segmentation method as in [26], [30]. A main contribution of this work is that we formulate patch-based label fusion in a probabilistic Bayesian framework.

1) The Bayesian Model: We will present the image registration method in the next section. Here, we assume that the transformation Φ_n between the target image I and the atlas image I_n has already been computed. We can then warp the atlas image and the corresponding label map to the target space. Let $I'_n(x) \equiv I_n(\Phi_n(x))$ denote the warped atlas image and $L'_n(x) \equiv L_n(\Phi_n(x))$ denote the warped label map.

In this model, we assume that each voxel x in the target image is generated from a corresponding voxel $x + \Delta x$ in one of the warped atlas images[3]. Δx denotes the registration error between the target image and the warped atlas image. We introduce a random vector field $M : \Omega \to \{1, \cdots, N\} \times \{1, \cdots, K\}$ to represent the mapping from each target voxel to its corresponding voxel in an atlas. N and K namely denote the number of atlases and the number of candidate voxels in a local neighbourhood. M is a vector field on the 3D lattice Ω. It is vector-valued and has the same size as the target image. At each voxel $x \in \Omega$, the first element of $M(x)$ maps this voxel x to an atlas, whereas the second element maps it to a candidate voxel. For instance, if we use a $3\times3\times3$ neighbourhood centred at x in the atlas image, there are 27 candidate voxels. Then $M(x) = [2, 5]$ denotes that the target voxel x corresponds to the 2nd atlas image and to the 5th candidate voxel in the neighbourhood around x. In Sabuncu's model, the mapping $M(x)$ for voxel x is a scalar, which denotes the index of an atlas [23]. In our model, we extend $M(x)$ to be a vector, which maps voxel x to a specific patch from an atlas. This feature enables us to account for the registration error and model the patch-based method in a Bayesian framework. This is the main difference between the two models.

The goal of segmentation is to estimate a label map L, given the image I and the atlas set $\{I'_n, L'_n|n = 1, \cdots, N\}$. It can be estimated by maximising a posteriori probability (MAP) of

[3]Note that the atlas image has already been warped into target space and the only difference comes from the registration error. Therefore, voxel x in the target image corresponds to voxel $x + \Delta x$ in the warped atlas image, instead of to $\Phi(x) + \Delta x$.

the label map conditioned on the target image and the atlas set,

$$\begin{aligned}\hat{L} &= \arg\max_{L} P(L|I,\{I'_n,L'_n\}) \\ &= \arg\max_{L} \frac{P(I,L|\{I'_n,L'_n\})}{P(I|\{I'_n,L'_n\})} \\ &= \arg\max_{L} P(I,L|\{I'_n,L'_n\}) \ , \end{aligned} \tag{1}$$

where $\hat{L}$ denotes the estimate of the label map L. The term $P(I|\{I'_n,L'_n\})$ is regarded as a constant and can be dropped in maximisation.

Introducing the mapping field M, the joint probability to observe the target image and label map $\{I,L\}$ conditioned on M can be written as,

$$\begin{aligned} & P(I,L|M,\{I'_n,L'_n\}) \\ =& \prod_{x\in\Omega} P(I(x),L(x)|M(x),\{I'_n,L'_n\}) \\ =& \prod_{x\in\Omega} P(I(x)|M(x),\{I'_n,L'_n\}) \cdot P(L(x)|M(x),\{I'_n,L'_n\}) \\ =& \prod_{x\in\Omega} P(I(x)|M(x),\{I'_n\}) \cdot P(L(x)|M(x),\{L'_n\}) \ . \end{aligned} \tag{2}$$

The first line above arises from the fact that once the mapping field M is known, each target voxel becomes conditionally independent, whose distribution is only determined by the atlas voxel that it maps to. Another effect of introducing M is that the intensity distribution and the label distribution becomes conditionally independent as well. The intensity distribution at a voxel is only dependent on the intensity of the atlas voxel that it maps to, whereas the label distribution at a voxel is also only dependent on the label map around that atlas voxel. Therefore, they are conditionally independent and the second and third lines of the equation arise. The conditional probabilities $P(I(x)|M(x),\{I'_n\})$ and $P(L(x)|M(x),\{L'_n\})$ will be analytically expressed in the following sections.

The probability that we want to maximise in Eq. (1) can then be calculated by marginalising the conditional probability w.r.t. M,

$$\begin{aligned} & P(I,L|\{I'_n,L'_n\}) \\ =& \sum_{M} P(M) \cdot P(I,L|M,\{I'_n,L'_n\}) \\ =& \sum_{M} P(M) \cdot \prod_{x\in\Omega} P(I(x)|M(x),\{I'_n\}) \cdot P(L(x)|M(x),\{L'_n\}) \end{aligned} \tag{3}$$

which is formed of the prior probability of the mapping field M, the conditional probability of the intensity $I(x)$ and the conditional probability of the label $L(x)$. We are going to describe the intensity likelihood and the probability likelihood in the next two subsections.

2) Intensity Likelihood: The mapping vector at voxel x is $M(x) = [M_1(x), M_2(x)]^T$, which means that voxel x in the target image is generated from the M_1-th atlas and from the M_2-th candidate voxel. We assume that the only difference between the target voxel and the atlas voxel is caused by Gaussian noise. Given $M(x)$, the conditional probability of intensity $I(x)$ can be written as,

$$\begin{aligned} & P(I(x)|M(x),\{I'_n\}) \\ =& \frac{1}{\sqrt{2\pi}\sigma_1} \exp\Big\{ -\frac{1}{2\sigma_1^2}\big[I(x) - I'_{M_1}(x+\Delta_{M_2})\big]^2 \Big\} \ , \end{aligned} \tag{4}$$

where Δ_{M_2} denotes the shift between the centre voxel x and the M_2-th candidate voxel, $I(x)$ and $I'_{M_1}(x+\Delta_{M_2})$ respectively denote the intensities at the target voxel and the corresponding atlas voxel, and finally σ_1 denotes the standard deviation of the Gaussian distribution.

The assumption of the Gaussian distribution is proposed due to its simplicity. It has to be noted that the Rician distribution is a more accurate model for intensity noise in MR images and the Gaussian distribution is a good approximation to the Rician distribution if the signal-to-noise ratio is high [45]. Also, since image interpolation is involved in the warping of the atlas image, the noise at each voxel is likely to be correlated. In addition, inhomogeneities and motion artefacts in cardiac MR images are not accounted here.

In order to yield a more robust estimate, we can compute the intensity likelihood based on a small patch instead of a single voxel[4],

$$\begin{aligned} & P(I(x)|M(x),\{I'_n\}) \\ \approx& \frac{1}{\sqrt{2\pi}\sigma_1} \exp\Big\{ -\frac{1}{2\sigma_1^2|S(x)|} \sum_{y\in S(x)} \big[I(y) - I'_{M_1}(y+\Delta_{M_2})\big]^2 \Big\}, \end{aligned} \tag{5}$$

where $S(x)$ denotes a patch centred at x, $|S(x)|$ denotes the number of voxels in $S(x)$. Here, the intensity difference of a single voxel is replaced by the mean squared difference in the patch $S(x)$. This is another difference of our model from Sabuncu's model in [23]. Because we assume that the intensity difference at a voxel, $[I(x) - I'_{M_1}(x+\Delta_{M_2})]$, follows a Gaussian distribution, the mean squared difference follows the $\chi^2(S(n))$ distribution with $S(n)$ degrees of freedom. According to the central limit theorem, it asymptotically converges to a normal distribution when $S(n)$ grows larger. Therefore, in Eq. (5), we approximately model the mean squared difference also using a normal distribution.

3) Label Likelihood: The label probability distribution models the potential registration error between the target image and the warped atlas. As a result, the target patch may correspond to an atlas patch which is possibly translated. For mathematical simplicity, we assume that registration error obeys normal distribution. Therefore, the conditional probability of label $L(x)$ can be written as,

$$\begin{aligned} & P(L(x)=l|M(x),\{L'_n\}) \\ =& \frac{1}{\sqrt{2\pi}\sigma_2} \exp\Big\{ -\frac{1}{2\sigma_2^2}|\Delta_{M_2}|^2 \Big\} \cdot \delta_{l,L_{M_1}(x+\Delta_{M_2})} \ , \end{aligned} \tag{6}$$

where $|\Delta_{M_2}|$ denotes the magnitude of the shift vector Δ_{M_2} between target patch and atlas patch, and $\delta_{i,j}$ denotes the Kronecker delta, which is equal to one when $i=j$ and equal to zero whenever $i \neq j$. The Kronecker delta is introduced so

[4]This is an engineering consideration. It is not mathematically strict.

that the label likelihood is non-zero if and only if this label l is equal to the label of the atlas voxel. The more distant the atlas patch is, the less impact it will have on the label.

Recently, several groups have proposed methods to estimate registration uncertainties within a Bayesian framework. Risholm et al. [46], [47] proposed to use Markov Chain Monte Carlo (MCMC) to estimate the posterior distribution of transformation, which allows non-parametric modelling of the registration uncertainty but can be computationally expensive. Simpson et al. [48], [49] proposed to infer an approximate posterior distribution of transformation parameters (for example, control point displacements for B-spline registration) using variational Bayes (VB). The distribution of the transformation parameters are assumed to be multivariate normal distribution. Since the displacement at each voxel is essentially a weighted sum of the transformation parameters within its support region, the registration uncertainty for the voxel can also be estimated and follows a normal distribution. In this work, we also assume that registration error obeys normal distribution[5].

4) Maximum A Posteriori (MAP): Let us assume that $M(x)$ is uniformly distributed for all $x \in \Omega$,

$$P(M(x)) = \frac{1}{N^{|\Omega|} \cdot K^{|\Omega|}},\ x \in \Omega.$$

As a result, the posterior probability in Eq. (3) can be maximised independently for each voxel,

$$\hat{L}(x) = \arg\max_{l} \sum_{M_1=1}^{N} \sum_{M_2=1}^{K} P(I(x)|M(x), \{I'_n\}) \cdot P(L(x) = l|M(x), \{L'_n\}) \ . \quad (7)$$

This can be viewed as weighted voting, where each of the K patches from each of the N atlases has a vote for the label l. Their votes are weighted according to the intensity similarity and the shift between the target patch and the atlas patch. The label with the highest vote is then selected. If $K = 1$, the method degenerates to a local-weighted label fusion model similar to the model in [20], where the registration error is not considered and only one patch is taken from each atlas.

This method is a local means method since the weight term is depending on the distance between the target patch and the atlas patch. However, if we do not weight the distance and disregard the label likelihood term in Eq. (7) [6], this method becomes a non-local means method and can be viewed as the patch-based segmentation method proposed in [26] or in [30]. In [26], the non-local means estimator is a binary case of Eq. (7). The label is either foreground or background and the target label is determined by thresholding the average label value by 0.5.

[5]Let $e_{reg,i}(x)$ denote the registration error at voxel x for atlas i and let $e_{seg,i}(x)$ denote the segmentation error at this voxel due to the registration error. In [31], [35], Wang et al. focuses on the segmentation error $e_{seg,i}(x)$ and estimates the optimal weights by minimising the total expected segmentation error, whereas our work focuses on the registration error $e_{reg,i}(x)$ and reduces weights for distant patches. Both works aim to account for the error in the warped atlas label map. However, the definition of the error term and the way to incorporate the error term is different.

[6]If $\sigma_2 \to +\infty$, the label likelihood in Eq. (6) tends to be a uniform distribution and can be disregarded.

In label fusion, we can generate a probabilistic label map $P_{\hat{L}}$ at the meantime,

$$P_{\hat{L}}(x) = [P_{\hat{L},1}(x), P_{\hat{L},2}(x), \cdots, P_{\hat{L},\mathcal{L}}(x)]^T \ , \quad (8)$$

where $P_{\hat{L}}(x)$ denotes the probabilistic vector at voxel x, the l-th element of the vector $P_{\hat{L},l}(x)$ denotes the probability to be the l-th tissue class and it is given by the summation in Eq. (7). The probabilistic label map will be used for image registration refinement in the next section.

C. Registration Refinement

In this section, we present the registration method which estimates the transformation between the target image and the atlas. Most existing methods consider registration separately from segmentation, where registration is solely based on information from intensity images. However, registration and segmentation are closely related and the solution of one can greatly assist in the computation of the other [50]–[53]. It has been shown that incorporating label information into the registration measure can significantly improve registration accuracy, compared to using intensity information only [52]. In this paper, we incorporate the consensus segmentation result into image registration in order to improve registration accuracy. With the improved registration, multi-atlas label fusion can be improved consequently. We name the strategy as "registration refinement". As far as we know, this is the first time that the consensus segmentation is fed back to image registration in a multi-atlas based approach. This is another contribution of this work.

The cost function for image registration is defined as,

$$\hat{\Phi}_n = \arg\max_{\Phi_n} f(I, I_n(\Phi_n)) + \omega \cdot g(P_{\hat{L}}, P_{L_n}(\Phi_n)) \ , \quad (9)$$

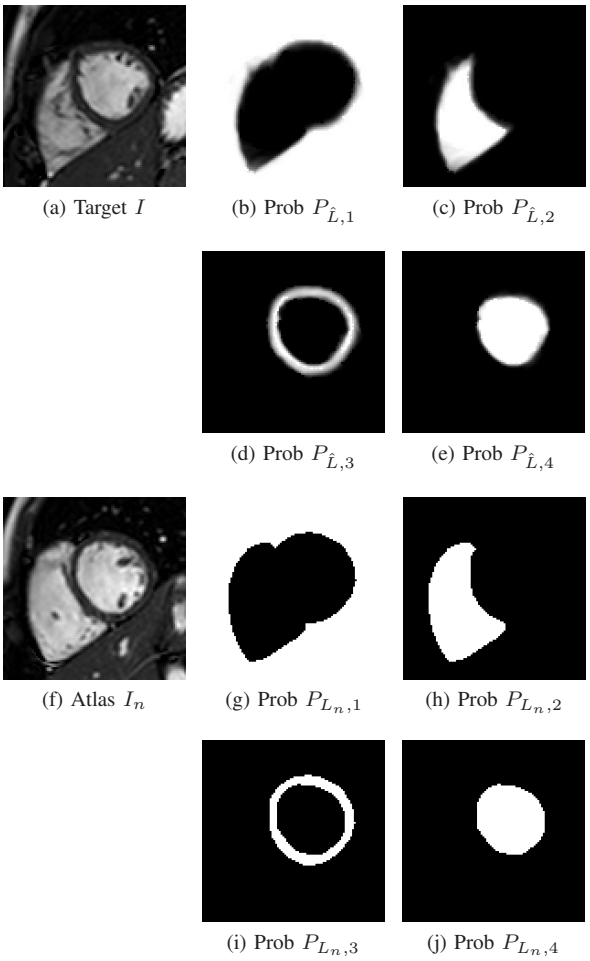
where Φ_n denotes the transformation between the target image and the n-th atlas. It is modelled by a free-form deformation based on B-splines [54]. $f(\cdot,\cdot)$ denotes the similarity metric between two intensity images, $g(\cdot,\cdot)$ denotes the similarity metric between two probabilistic label maps and ω balances the weight. Figure 2 illustrates the intensity images and the probabilistic label maps respectively for target and atlas.

We use normalised mutual information (NMI) as the similarity metric for the intensity images, which is widely used in the image registration domain and is efficient to calculate [55]. We also tried mean squared difference (MSD) as the similarity metric for image registration and found that NMI performed slightly better.

The probabilistic label map for the target image $P_{\hat{L}}$ is computed when weighted label fusion is performed, whereas the probabilistic label map for the n-th atlas P_{L_n} is produced from the discrete label map L_n,

$$P_{L_n}(x) = [\delta_{1,L_n(x)},\ \delta_{2,L_n(x)},\ \cdots,\ \delta_{\mathcal{L},L_n(x)}]^T \ . \quad (10)$$

The probabilistic vector $P_{L_n}(x)$ takes discrete values. However, since interpolation is involved, the transformed vector $P_{L_n}(\Phi(x))$ will take continuous values. Linear interpolation is used because it is fast to compute.



(a) Target I (b) Prob $P_{\hat{L},1}$ (c) Prob $P_{\hat{L},2}$

(d) Prob $P_{\hat{L},3}$ (e) Prob $P_{\hat{L},4}$

(f) Atlas I_n (g) Prob $P_{L_n,1}$ (h) Prob $P_{L_n,2}$

(i) Prob $P_{L_n,3}$ (j) Prob $P_{L_n,4}$

Fig. 2: The intensity images and the probabilistic label maps for target and atlas. The probabilistic label map is a vector-valued image. The four components of the probability vector are namely background, RV, myocardium and LV. The target probabilistic label map is estimated during label fusion, whereas the atlas probabilistic label map is produced from manual labelling. Image registration will use both intensity information and label information.

The similarity metric between the probabilistic maps $P_{\hat{L}}$ and $P_{L_n}(\Phi_n)$ is defined as,

$$g(P_{\hat{L}}, P_{L_n}(\Phi_n)) = \sum_{x\in\Omega} \langle P_{\hat{L}}(x), P_{L_n}(\Phi_n(x)) \rangle \ , \qquad (11)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors. We name this metric as *probabilistic label consistency*. When two probabilistic maps exactly match, the probabilistic label consistency yields one. This metric is a continuous version of the *label consistency* metric proposed in [56], which is defined as an overlap metric between two discrete label maps.

An alternative is to generate a discrete label map after label fusion from the hard segmentation and use the discrete label consistency metric in the cost function Eq. (9). The reason why we use the probabilistic label consistency instead of the discrete one is that we want to preserve the probabilistic information after label fusion, which is a soft segmentation. It has to be noted that although the label fusion model in the previous section is derived in a Bayesian framework, the image registration part is defined in a more ad-hoc way.

The rationale behind the proposed cost function Eq. (9) is that if the target image and the atlas image share the same anatomical structure and topology, and the ground truth label maps for them are both known, there should be an optimal transformation Φ, which simultaneously maps the target intensity image I to atlas intensity image I_n and the target label map P_L to atlas label map P_{L_n}, i.e.

$$\hat{\Phi}_n = \arg\max_{\Phi_n} f(I, I_n(\Phi_n)) + \omega \cdot g(P_L, P_{L_n}(\Phi_n)) \ . \quad (12)$$

Because labels can be less ambiguous than intensity values in differentiating tissue classes (for example, the same intensity value in an image may correspond to different classes but the same label will not), the incorporation of label information into the cost function is likely to result in a more accurate registration between the two anatomies. However, since the target label map P_L is still unknown, we approximate it using the estimated label map $P_{\hat{L}}$ and substitute Eq. (12) by Eq. (9) in the optimisation. Since the estimate $P_{\hat{L}}$ comes from multi-atlas label fusion, it may be a good approximation to the ground truth P_L, if the majority of the target-to-atlas registrations are good.

Image registration is initialised by using intensity images only with $\omega = 0$. Using the initial registration results, label fusion is performed and the target probabilistic label map $P_{\hat{L}}$ is obtained. Then, image registration will use both intensity images and probabilistic label maps in order to improve registration accuracy. We refer to this as "registration refinement". Afterwards, the method can iterate between label fusion and registration refinement until convergence (for example, when the change of the label map estimate becomes small):

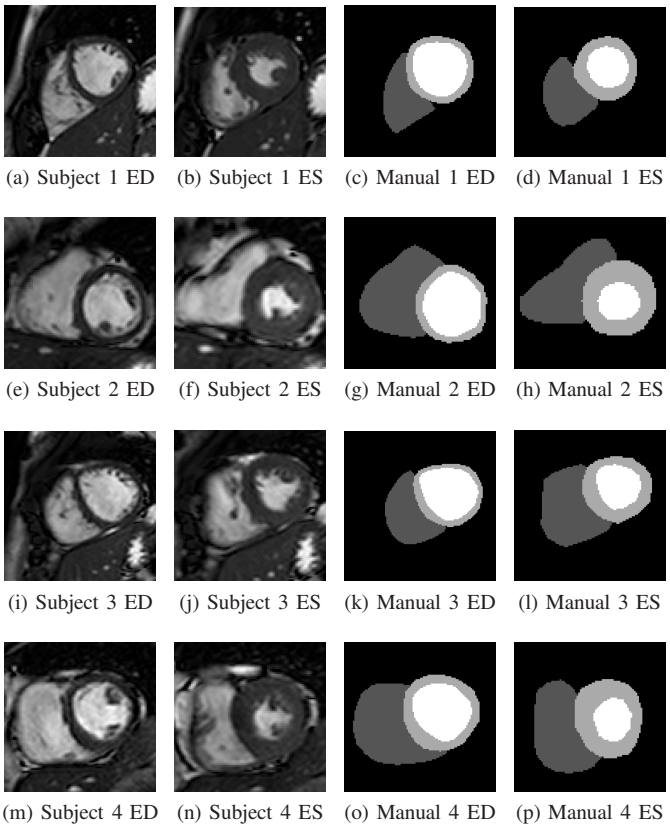
1) Label fusion is performed, using the registration results to warp the atlases.
2) Registration refinement is performed, using the target probabilistic label map estimated from previous label fusion as input.

The two steps respectively update label fusion using more accurate registration and then update registration using more accurate label information, in the hope that more accurate registration can result in a more accurate label estimate, and vice versa. In the following experiments, however, we will show that after just one iteration, segmentation performance is already significantly increased.

III. Experiments and Results

A. Data

In order to evaluate the performance of the proposed segmentation method, we tested it on a short-axis cardiac MR (CMR) data set of 28 subjects using leave-one-out cross validation. When each image is segmented, the other 27 images with corresponding label maps are regarded as the atlas set. The CMR data were acquired on a 1.5T Philips Achieva system (Best, Netherlands) using a 32-channel coil and the balanced-steady state free precession (b-SSFP) sequence. Images in the left ventricular short-axis plane were acquired using the following parameters: field-of-view, 320×320mm; repetition time (TR), 3.0ms; echo time (TE), 1.5ms; shot duration,



(a) Subject 1 ED (b) Subject 1 ES (c) Manual 1 ED (d) Manual 1 ES

(e) Subject 2 ED (f) Subject 2 ES (g) Manual 2 ED (h) Manual 2 ES

(i) Subject 3 ED (j) Subject 3 ES (k) Manual 3 ED (l) Manual 3 ES

(m) Subject 4 ED (n) Subject 4 ES (o) Manual 4 ED (p) Manual 4 ES

Fig. 3: Some examples of the testing images and corresponding manual labelling at end-diastole (ED) and end-systole (ES). As can be seen, the cardiac anatomical shape is highly variable across subjects.

50ms; number of cardiac phases, 30; section thickness, 8mm with a 2mm gap. The reconstructed MR images are of dimension 288×288×12, with voxel spacing 1.23×1.23×10mm. The LV cavity, LV myocardium and the RV cavity were manually labelled by two experienced imaging scientists for both ED and ES frames. 10 subjects were labelled by one observer, whereas the other 18 were labelled by the other observer. Each subject was labelled exactly once. Figure 3 displays the images and corresponding manual labellings for some example subjects. As can be seen, the cardiac anatomical shape is highly variable across subjects.

B. Parameter Settings

Six subjects were randomly selected. The diastolic frames of these subjects were used as the training set for parameter tuning. First we tuned the parameters for label fusion, without applying the registration refinement. $\sigma_1 = 50$, $\sigma_2 = 1.5$mm and a 2D patch size of 4×4mm (3×3×1voxels) were found to perform well on the training set. We also tested a 3D patch size of 4×4×30mm (3×3×3voxels) and found that the 2D patch size performed better. This is probably because the 2D patch is large enough to capture the local geometry but not too large to lose localisation, whereas the 3D patch is too elongated due to the large slice thickness of 10mm. The neighbourhood for the atlas patches were set to $\pm 2\sigma_2$ around the centre voxel, since we assume a Gaussian distribution for the registration error and it decays to negligible levels after $2\sigma_2$.

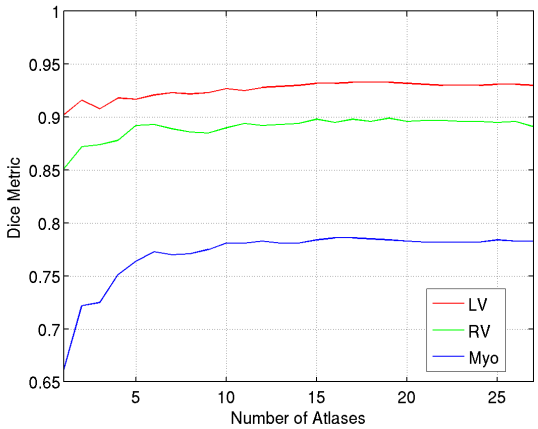


Fig. 4: Effect of the size of atlas subset on segmentation accuracy. The accuracy is evaluated using the Dice overlap metric between automatic segmentation results and manual label maps for the training set. The Dice metric is displayed for the LV cavity, myocardium and the RV cavity respectively.

In order to save computation and discard dissimilar atlases in label fusion, we use the atlas selection strategy as proposed in [17]. All the atlases are first registered to the target image using affine transformation, which is computationally fast. After affine registration, only a subset of atlases that are most similar to the target image in terms of normalised mutual information (NMI) are selected and used in the next stage of non-rigid registration and label fusion, since non-rigid registration is computationally more expensive. Atlas selection also has the benefit that some very dissimilar atlases can be discarded before label fusion, therefore segmentation accuracy is not affected by these atlases. Figure 4 shows the effect of the size of atlas subset on segmentation accuracy. It shows that for the training set, the segmentation performance climbs up as more atlases are used until it gradually levels. After 15 atlases, using more atlases results in only trivial improvement or even slight decline of performance, because atlases dissimilar to the target image begin to be introduced. Aljabar et al. reported a similar trend in [17]. In the following experiments, we will select 15 atlases for label fusion.

After we fixed the parameters for label fusion, we tuned the parameters for registration refinement, including the weight term ω and the number of iterations. $\omega = 1.5$ was found to perform well. The number of iterations for registration refinement was set to 1, because we found that the improvement due to introducing the label information is most significant after the first iteration but becomes less pronounced during further iterations, as shown in Figure 5. The reason is that at the first iteration, we start to introduce label information into the registration so the cost function has changed considerably. Afterwards, the change of the cost function is only due to an updated label map which is less drastic. Each iteration involves re-running registration for all the selected atlases, which is computationally expensive. Therefore, one iteration is a good compromise which improves segmentation accuracy without increasing the computational burden too much.

We have also found that after about 5 iterations, the algo-

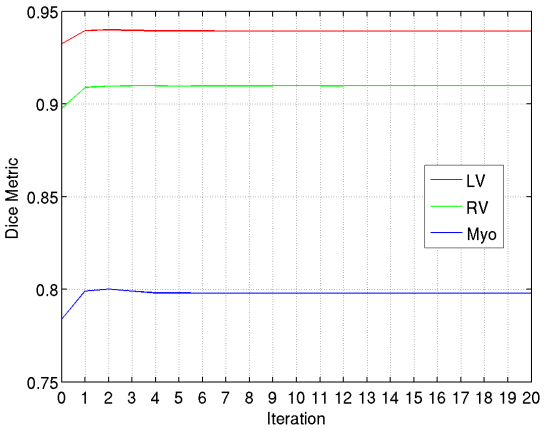


Fig. 5: Effect of the iterative registration refinement on segmentation accuracy evaluated using the Dice metric, using $\omega = 1.5$. Iteration 0 denotes the initial non-rigid registration using intensity images only. Improvement due to introducing the label information is most significant after the first iteration but becomes less pronounced during further iterations.

rithm starts to converge. The reason is that as the change of the label map estimate becomes small, the deformation field Φ only changes little during image registration. After a few iterations, it even stops changing because the change of the registration cost function is so small that the gradient descent algorithm converges to a local solution. As a result, in the next update of the label estimate, the label estimate $P_{\hat{L}}$ also stops changing. However, it is difficult to obtain an analytical proof of the convergence and its properties due to the complexity of this interleaved problem.

TABLE I: Comparison of the Dice overlap metric for a combination of different registration and label fusion strategies. The bold text denotes the highest value in each table. NRR stands for non-rigid registration with refinement.

(a) Left Ventricle

	Registration Strategy		
Fusion Strategy	Affine	Non-Rigid	NRR
Majority voting	0.846	0.905	0.915
Local-weighted	0.879	0.907	0.914
Patch-based	0.887	0.910	**0.915**

(b) Right Ventricle

	Registration Strategy		
Fusion Strategy	Affine	Non-Rigid	NRR
Majority voting	0.811	0.870	0.884
Local-weighted	0.841	0.871	0.885
Patch-based	0.850	0.874	**0.886**

(c) Myocardium

	Registration Strategy		
Fusion Strategy	Affine	Non-Rigid	NRR
Majority voting	0.662	0.800	0.818
Local-weighted	0.736	0.807	0.819
Patch-based	0.752	0.814	**0.824**

C. Segmentation Accuracy

We performed segmentation for all the 28 subjects using a combination of different registration and label fusion strategies. Registration strategies include affine registration (Affine), non-rigid registration (NonRigid) and non-rigid registration with refinement (NRR). Label fusion strategies include majority voting (MV), local-weighted label fusion (LW) and patch-based label fusion (PB). By "local-weighted label fusion", we mean that in Equation (7), $K = 1$ and the weight is determined only by the similarity between the target patch and a single atlas patch. By "patch-based label fusion", we mean that multiple atlas patches in a neighbourhood are considered in order to account for the registration error. In the following comparison, we will name each method by concatenating the registration and label fusion names. For example, "NRR-PB" denotes using non-rigid registration with refinement (NRR) and patch-based label fusion (PB).

The Dice overlap metric between the automatic segmentation and the manual labelling is computed to evaluate segmentation accuracy. Table I lists the mean Dice metric for different registration and label fusion strategies. The Dice metric is averaged over the 28 subjects and over the ED and ES frames. As we can see from the table, NRR-PB consistently provides the highest Dice metric for all the three tissue classes, namely the left ventricle, right ventricle and myocardium, because this method not only refines registration accuracy but also accounts for potential registration errors in label fusion. Affine-MV results in the lowest Dice metric, because affine registration may not be very accurate and majority voting does not consider the intensity similarity and the registration error between target and atlas.

When we compare methods using the same label fusion strategy (each row in the table), more accurate registration results in higher segmentation accuracy. For example, when PB is used for label fusion, the mean Dice metric of the LV is 0.915 for NRR-PB, which is significantly higher than 0.887 for Affine-PB ($p < 0.001$) and 0.910 for NonRigid-PB ($p < 0.001$). When we compare methods using the same registration strategy (each column in the table), the segmentation accuracy increases as the label fusion strategy becomes more sophisticated. For example, when affine registration is used, the Dice metric of the LV is 0.846 for Affine-MV, 0.879 for Affine-LW and 0.887 for Affine-PB. The Dice metric of Affine-PB is significantly higher than Affine-MV ($p < 0.001$) and Affine-LW ($p < 0.001$).

We have also found that when registration becomes very accurate, different label fusion strategies do not make a big difference in segmentation performance. When NRR is used, the Dice metrics for MV, LW and PB become very close, for example, namely 0.915, 0.914 and 0.915 for the LV, without statistical difference ($p > 0.1$). When the registration is less accurate, however, the difference between different label fusion strategies becomes more evident. For example, when affine registration is used, Affine-PB significantly outperforms Affine-MV, especially for the myocardium ($p < 0.001$).

In a recent paper on cardiac image segmentation [8], Zhuang et al used a locally affine registration method (LARM) to propagate a single average atlas to the target image. They reported that the mean Dice metric was 0.92 for the LV, 0.87 for the RV and 0.77 for the myocardium. Our results are comparable to theirs for the LV, slightly better for the RV

and also better for the myocardium. It has to be noted that Zhuang's experiments were based on a different image set of $1\times1\times1$mm isotropic resolution, whereas our images are short-axis stacks with a much lower anisotropic resolution of $1.23\times1.23\times10$mm. However, Zhuang et al performed whole heart segmentation (four chambers and myocardium) by using a detailed whole heart atlas, whereas we only performed segmentation for the LV, RV and myocardium.

TABLE II: Comparison of the mean distance error (unit: mm) for a combination of different registration and label fusion strategies. The bold text denotes the smallest distance in each table. NRR stands for non-rigid registration with refinement.

(a) Left Ventricle Endocardium

	Registration Strategy		
Fusion Strategy	Affine	Non-Rigid	NRR
Majority voting	2.44	1.39	1.29
Local-weighted	1.72	1.34	1.28
Patch-based	1.62	1.31	**1.26**

(b) Left Ventricle Epicardium

	Registration Strategy		
Fusion Strategy	Affine	Non-Rigid	NRR
Majority voting	2.65	1.68	1.49
Local-weighted	2.37	1.63	**1.48**
Patch-based	2.34	1.63	1.49

(c) Right Ventricle Endocardium

	Registration Strategy		
Fusion Strategy	Affine	Non-Rigid	NRR
Majority voting	2.97	1.94	1.70
Local-weighted	2.25	1.84	1.69
Patch-based	2.19	1.81	**1.68**

TABLE III: Comparison of the maximum distance error (unit: mm) for a combination of different registration and label fusion strategies. The bold text denotes the smallest distance in each table. NRR stands for non-rigid registration with refinement.

(a) Left Ventricle Endocardium

	Registration Strategy		
Fusion Strategy	Affine	Non-Rigid	NRR
Majority voting	10.16	8.41	7.27
Local-weighted	10.24	8.49	7.32
Patch-based	9.83	7.99	**7.27**

(b) Left Ventricle Epicardium

	Registration Strategy		
Fusion Strategy	Affine	Non-Rigid	NRR
Majority voting	12.66	10.54	9.49
Local-weighted	13.30	11.48	9.47
Patch-based	13.25	10.94	**9.35**

(c) Right Ventricle Endocardium

	Registration Strategy		
Fusion Strategy	Affine	Non-Rigid	NRR
Majority voting	14.00	12.74	12.18
Local-weighted	14.31	13.20	**12.15**
Patch-based	14.11	12.98	12.23

Tables II and III respectively lists the mean and the maximum distance error (the Hausdorff distance) between the manually delineated surface and the surface given by our segmentation for different registration and label fusion strategies. In terms of the mean distance error, when the same registration strategy is used, the patch-based label fusion almost always outperforms the other two. For example, in Table II, the mean distance error of the LV endocardium is 1.62 mm for Affine-PB, which is significantly lower than 1.72 mm for Affine-LW ($p < 0.001$) and 2.44 mm for Affine-MV ($p < 0.001$). In addition, NRR always improves the performance compared to the other two registration strategies. For example, the mean distance error of the LV endocardium is 1.26 mm for NRR-PB, which is significantly lower than 1.31 mm for NonRigid-PB ($p < 0.01$) and 1.62 mm for Affine-PB ($p < 0.001$).

The maximum statistic is more sensitive to outliers than the mean. Looking at the maximum distance error in Table III, we have observed that sometimes the patch-based label fusion may not result in the lowest distance error when the same registration is used. However, when the same label fusion strategy is used, NRR still always outperforms the other two registrations. For example, the maximum distance error of the LV endocardium is 7.27 mm for NRR-PB, significantly lower than 7.99 mm for NonRigid-PB ($p < 0.01$) and 9.83 mm for Affine-PB ($p < 0.001$).

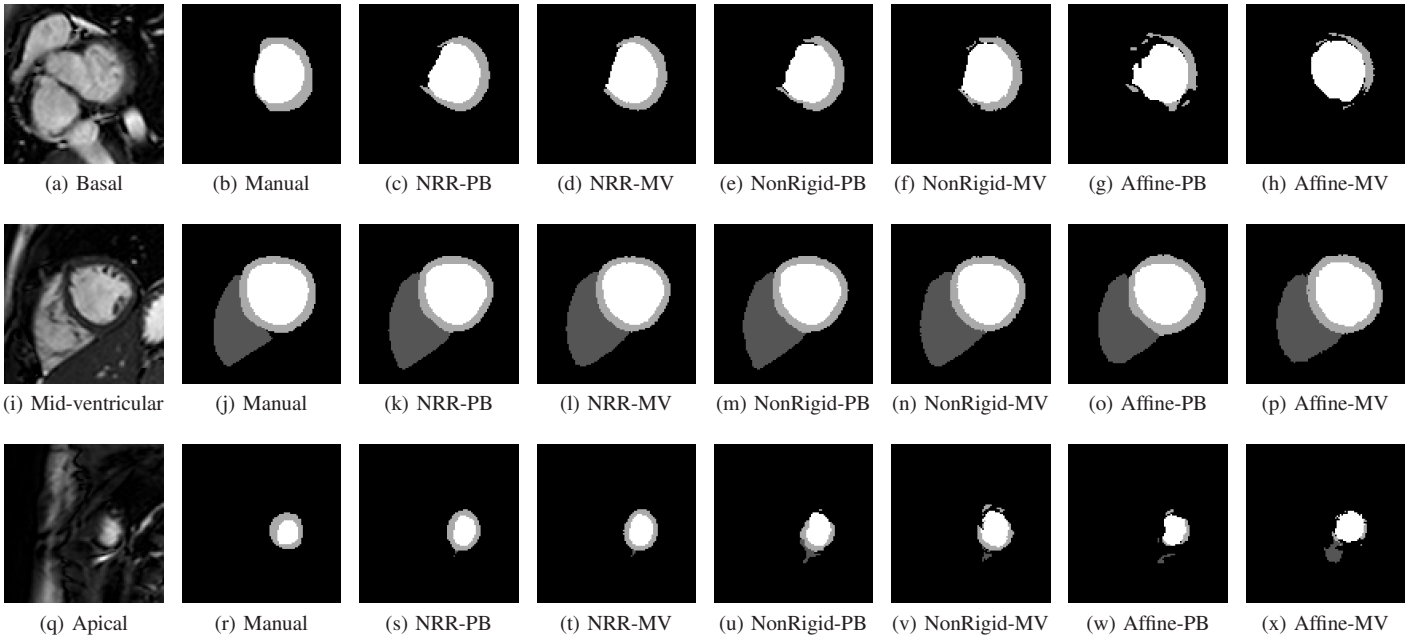
Figure 6 displays an example of the segmentation results for different registration and label fusion strategies. It shows that both NRR-PB and NRR-MV perform well in all the three slices. NonRigid-PB performs well on the mid-ventricular slice. However, on the apical slice, it produces broken myocardium segmentation. Similarly, NonRigid-MV performs well on the mid-ventricular slice but produces disconnected label maps on the basal and apical slices. Affine-PB and Affine-MV perform even worse on the basal and apical slices. We have found that all the methods perform well on the mid-ventricular slice because on this slice the contrast between the myocardium, the blood pool and the background is very strong. The apical slice is more difficult to segment, since the contrast of the myocardium becomes low.

TABLE IV: Comparison of the computation time for different registration strategies.

	Affine	Non-Rigid	NRR
Per Atlas	15 sec	2 min	5 min

D. Computation Time

Table IV lists the computation time per atlas for each image registration strategy. It took NRR about 5 minutes to perform registration for one atlas, which is a little over twice the time as for NonRigid. This is because NRR evaluates two terms in the registration cost function, one for intensity information and the other for label information, whereas NonRigid only evaluates the intensity term. Registrations for all the atlases were performed parallel on a 32-core computing server. Therefore, it took about 5 minutes for NRR to finish all the registrations for 15 pre-selected atlases. Considering there are more and more affordable desktop computers equipped with 8 or 16 CPU cores these days, performing multi-atlas segmentation for clinical applications becomes practical in terms of time. Although NonRigid-PB takes less time, NRR-PB can lead



(a) Basal (b) Manual (c) NRR-PB (d) NRR-MV (e) NonRigid-PB (f) NonRigid-MV (g) Affine-PB (h) Affine-MV

(i) Mid-ventricular (j) Manual (k) NRR-PB (l) NRR-MV (m) NonRigid-PB (n) NonRigid-MV (o) Affine-PB (p) Affine-MV

(q) Apical (r) Manual (s) NRR-PB (t) NRR-MV (u) NonRigid-PB (v) NonRigid-MV (w) Affine-PB (x) Affine-MV

Fig. 6: Comparison of the segmentation results for different registration and label fusion strategies. Top to bottom: the basal, mid-ventricular and apical slices from one of the subjects. Left to right: original images, manual labelling, NRR-PB, NRR-MV, NonRigid-PB, NonRigid-MV, Affine-PB and Affine-MV.

to a more accurate segmentation with consistent contours, as shown in Figure 6. This will be beneficial not only for the measurement of clinical indices such as EDV and ESV, but also for building a patient-specific cardiac mesh model which normally requires consistent contours in the segmentation.

Regarding the computation time for label fusion, MV, LW and PB respectively took about 0.1, 0.5 and 5 seconds. This is negligible compared to the time used for image registration.

E. Clinical Indices

We measured the end-diastolic volume (EDV) and end-systolic volume (ESV) for the left ventricle based on the segmentation. The ejection fraction (EF) was then calculated using the following equation,

$$\mathrm{EF} = \frac{\mathrm{EDV} - \mathrm{ESV}}{\mathrm{EDV}} \times 100\% \ . \tag{13}$$

The ventricular mass (VM) was calculated from the end-diastolic myocardial volume and using 1.05 gram/cm^3 as the density [57].

Figure 7 shows the Bland-Altman plots for segmentation results using NRR-PB. The difference between automatic measurement and ground truth is plotted against the ground truth. The Bland-Altman plots show that for most of the subjects, the automated measurement is in good agreement with the manual measurement. Our segmentation method can produce relatively accurate estimation of clinical indices.

TABLE V: Comparison of the mean Dice metrics between our method and its hierarchical version. Hier stands for hierarchical weighted label fusion.

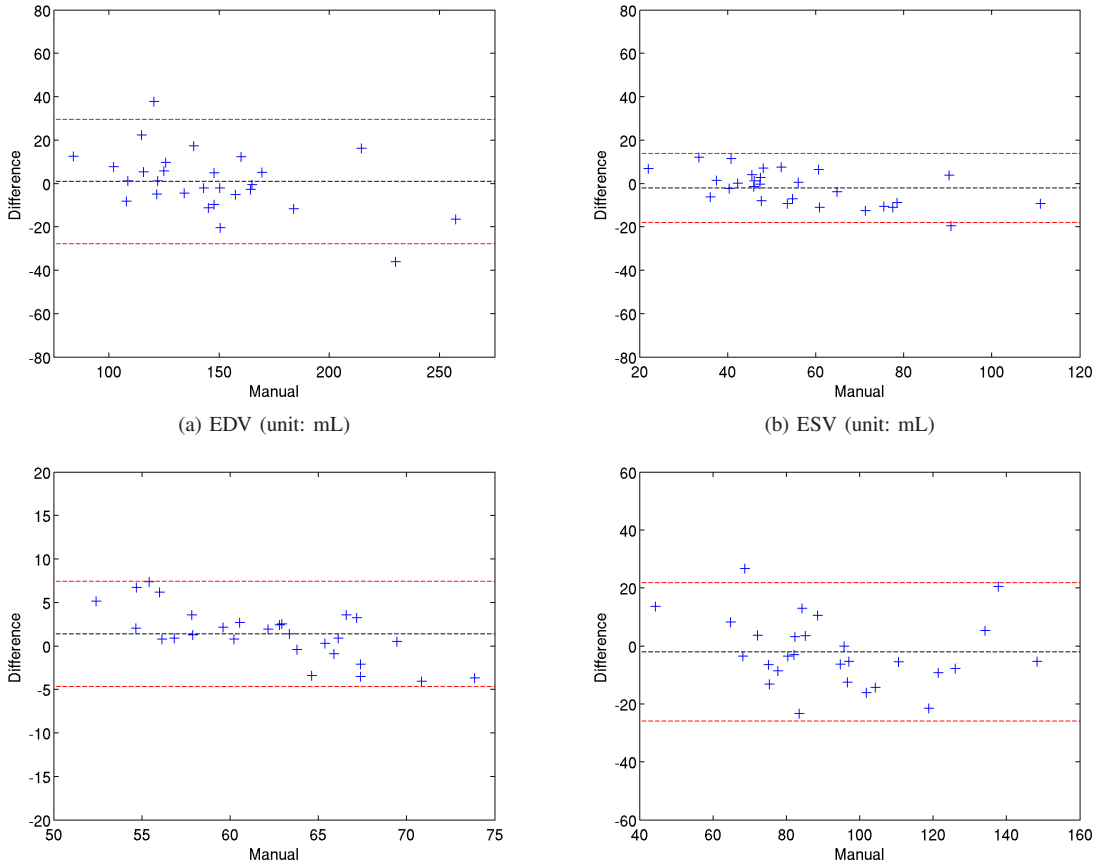
	NRR-PB	NonRigid-Hier	NRR-Hier
Left ventricle	0.915	0.908	0.915
Right ventricle	0.886	0.871	0.883
Myocardium	0.824	0.811	0.823

F. Comparison to Hierarchically Weighted Label Fusion

Wolz et al. [38] proposed to apply patch-based segmentation to abdominal CT scans with a hierarchical weighting scheme, which consists of weighting at global level, organ level and voxel level. Our method already consists of the global level (atlas selection) and voxel level (patch-based similarity) weighting. For comparison to Wolz's method, we add an organ level weighting into our method, which evaluates the weight based on the organ-wise similarity between target and atlas as well as the organ-wise agreement between atlases [38].

Table V compares the segmentation performance of our method (NRR-PB) to its hierarchical version, combined with both non-rigid registration (NonRigid-Hier) and registration refinement (NRR-Hier). The original paper of Wolz et al. used non-rigid registration only and corresponds to NonRigid-Hier in this table. We have found that our method (NRR-PB) performs significantly better than NonRigid-Hier ($p < 0.001$ for LV, RV and Myo) and slightly better than NRR-Hier ($p < 0.01$ for RV and $p > 0.1$ for LV and Myo).

It has to be noted that Wolz's hierarchical method was designed for multi-organ segmentation in abdominal CT scans, where the liver, spleen, pancreas and kidney are distributed across a large field-of-view. Therefore, registration is more difficult and each pairwise target-to-atlas registration can perform differently for different organs, for example, achieving a good registration for the liver but a very bad registration for the kidney and the target kidney is not aligned with the atlas kidney at all. In that case, only using patch-based similarity may associate the target patch with atlases patches with a similar intensity pattern but very far from the kidney, even from another organ. As a result, the target patch is likely to be wrongly labelled. To counteract this effect, an organ-wise weight was introduced to evaluate the organ-wise registration performance and to assign a higher weight to organs that are



(a) EDV (unit: mL)

(b) ESV (unit: mL)

(c) Ejection fraction (unit: %)

(d) Ventricular mass (unit: g)

Fig. 7: The Bland-Altman plots for segmentation results using NRR-PB. The difference between automatic measurement and ground truth is plotted against the ground truth. The dark dashed line denotes the mean difference. The red dashed lines denote ± 2 standard deviation.

well registered. In our application of cardiac images, the LV, RV and myocardium are located closely to each other in a small field-of-view. The image registration is unlikely to align the heart with another organ. Therefore, an additional organ-wise weighting may not be so important.

TABLE VI: Comparison of the mean Dice metrics between our method and the sparse method.

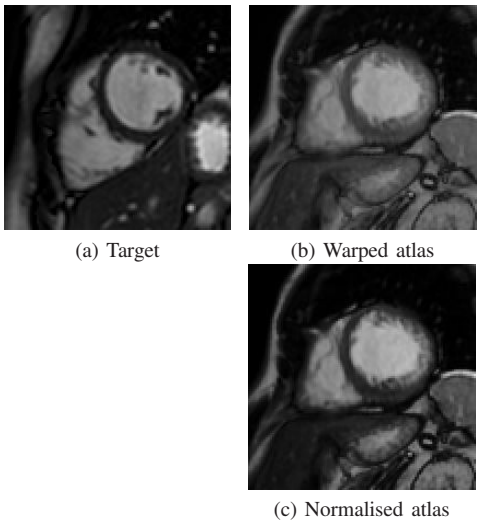
	NRR-PB	NRR-Sparse	Affine-PB	Affine-Sparse
Left ventricle	0.915	0.909	0.887	0.881
Right ventricle	0.886	0.883	0.850	0.836
Myocardium	0.824	0.814	0.752	0.766

G. Comparison to Sparse Representation Classification

Tong et al. [32] proposed a sparse patch representation method for hippocampus segmentation in brain images. The atlas patches are considered as atoms in a dictionary and each target patch is represented by a sparse linear combination of the atoms (patches). The target label is then determined by very few sparsely selected atlas patches, instead of by all the atlas patches in the neighbourhood. Here we also compare the proposed method to the sparse representation method. Table VI lists the mean Dice metric for our method and the sparse method. The original paper of Tong et al. used affine registration only and corresponds to Affine-Sparse in this table. We have found that our method (NRR-PB) performs significantly better than Affine-Sparse ($p < 0.001$ for LV, RV and Myo) and also slightly better than NRR-Sparse ($p < 0.001$ for LV and Myo and $p > 0.01$ for RV). When we compare Affine-PB to Affine-Sparse, however, we have found that Affine-Sparse performs significantly better than Affine-PB for the myocardium ($p < 0.01$). This is probably because when registration becomes less accurate such as when using affine registration, there are fewer atlas patches in the local search neighbourhood which are of the same tissue class as the target patch especially for the myocardium which is relatively thin and therefore the sparse representation of the target patch becomes reasonable. This is similar to the case in Tong's work [32], where the hippocampus is relatively a small structure and there are far more background patches than hippocampus patches. In this kind of situation, sparse representation becomes a powerful tool to pick up the hippocampus patches.

H. Using Atlases from a Different Hospital

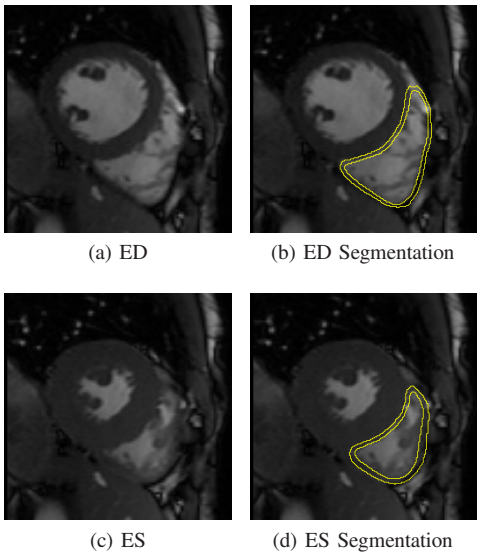
In the proposed label fusion model, we use the mean squared difference metric and a Gaussian model to weight the patch similarity. This model works well if all the images are acquired at the same site, the same scanner and using the same acquisition protocol so that the intensity scales are similar. If the atlas image have a different intensity scale from the target



(a) Target (b) Warped atlas

(c) Normalised atlas

Fig. 8: Segmentation using atlases acquired from a different hospital. (a) Target image from Set 1; (b) Warped atlas image from Set 2; (c) Atlas image after intensity normalisation.



(a) ED (b) ED Segmentation

(c) ES (d) ES Segmentation

Fig. 9: RV endocardium and epicardium segmentation respectively for the ED and ES phase using NRR-PB.

image, we may need to perform intensity normalisation prior to the calculation of the mean squared difference metric.

In this experiment, we segmented the 28 cardiac MR images (Set 1) involved in the previous experiments using a different set of 30 atlases (Set 2) acquired at a different hospital with different acquisition parameters. Set 2 was acquired on a 1.5T Philips Achieva system (Best, Netherlands) with the SSFP sequence using the following parameters: TR 2.9ms; TE 1.5 ms; volume dimension 256×256×12; voxel spacing 1.45×1.45×10mm. Apart from these parameters, another major difference is that the subjects in Set 2 are patients, whereas the subjects in Set 1 are normal volunteers. As a result, the image quality in Set 2 is sometimes worse than Set 1 since some patients could not hold their breath steadily during the acquisition. The LV cavity and LV myocardium were labelled by an experienced imaging scientist for the ED frame for Set 2. Therefore, we only segmented the LV and myocardium at the ED frame on Set 1 using the new atlas set.

Before label fusion, intensity normalisation was applied to the warped atlas images from Set 2 by matching its intensity histogram to the target image histogram [58]. Figures 8 (a) and (b) respectively shows a target image from Set 1 and a warped atlas image from Set 2. As can be seen, the myocardium displays different intensity scales in the two images. However, after intensity normalisation, Figure 8 (c) shows an intensity scale similar to that of the target image, so that patch comparison between the two images becomes possible.

TABLE VII: Comparison of the mean Dice overlap metric, using selected atlases from different atlas sets.

	15 atlases from Set 1	15 atlases from Set 2	30 atlases from Sets 1+2
Left ventricle	0.940	0.924	0.941
Myocardium	0.799	0.754	0.802

Table VII lists the mean Dice overlap metrics for ED frame segmentation over the 28 subjects of Set 1 using NRR-PB. We compare the segmentation performance using three different atlas sets, namely 15 atlases selected from Set 1, 15 atlases selected from Set 2 and finally 30 atlases selected from Sets 1 and 2 combined. The atlas selection method is the same as described in Section III-B. As Table VII shows, when atlases from Set 2 are used to segment images from Set 1, we can still achieve a Dice metric of 0.924 for the LV and 0.754 for the myocardium, though it is lower than using atlases from Set 1 itself ($p < 0.001$). If we combine Sets 1 and 2 and select 30 atlases from the combined set, we can achieve a Dice metric of 0.941 for the LV and 0.802 for the myocardium, which is slightly higher than using 15 atlases from Set 1 but without statistical significance ($p > 0.1$). However, this experiment still shows that atlases acquired with a different scanner and a different protocol can be used for the proposed method after performing intensity normalisation for the atlases.

TABLE VIII: The mean Dice overlap metric (unit:1) and the mean Hausdorff distance (HD, unit:mm) for RV endocardium and epicardium on the MICCAI RV Segmentation Challenge Data.

		NonRigid-PB	NRR-PB
Dice	RV endo	0.755	0.794
	RV epi	0.799	0.829
HD	RV endo	9.77	8.63
	RV epi	10.34	9.26

I. Application to MICCAI RV Segmentation Challenge Data

We participated in the MICCAI 2012 RV Segmentation Challenge [11], [12], which was an on-site competition for the segmentation of the RV endocardium and epicardium, and won the first prize in the category for semi-automatic algorithms[7]. Figure 9 shows an example of the segmentation results using NRR-PB. As it shows, although the RV myocardium is thin and a little vague, it is well delineated by our segmentation. Table VIII lists the results for NonRigid-PB and NRR-PB. For the on-site competition, we used the NonRigid-PB method.

[7]Five landmarks are used to initialise the RV image registration, which takes only about 10 seconds per subject for human intervention. All the other steps are automatic.

IV. Discussion and Conclusions

In the experiments, we have found that registration accuracy has a great impact on segmentation performance. With registration refinement, label fusion results can be improved. In addition, when registration refinement is used, the difference between different label fusion strategies becomes subtle. Even majority voting can perform quite well in this case. However, if the registration is not very accurate, for example when affine registration is used, sophisticated label fusion strategies such as the patch-based method play a very important role in improving segmentation performance.

In Sabuncu's model [23], the mapping field M was assumed as a Markov random field (MRF). Three different cases were discussed, namely local, global and semi-local weighted voting. The latter two introduce regularisation for the mapping field in order to smooth the resulting label map. In this paper, the label map is not explicitly regularised. However, since the weights are calculated from patches, it has an effect of intrinsic regularisation [20]. For instance, the patches around two neighbouring voxels in the target image have a large percentage of overlap. When the weights between these two patches and an atlas patch are calculated, the overlap can result in smooth variation of the weights and subsequently smooth variation of the label map estimate.

Multi-atlas label fusion involves estimation for the label map L and estimation for the deformation fields $\{\Phi_n\}$. The proposed method can be regarded as an engineering solution to the following optimisation problem:

$$\hat{L}, \{\hat{\Phi}_n\} = \arg\max_{L,\{\Phi_n\}} P(L|I, \{I_n, L_n, \Phi_n\}) + \sum_n [f(I, I_n(\Phi_n)) + \omega \cdot g(P_{\hat{L}}, P_{L_n}(\Phi_n))] \ . \quad (14)$$

In an alternating optimisation manner, we alternatively fix Φ_n and update the first term in the cost function related to label fusion, then fix L and update the remaining terms related to image registration, so that the decomposed problems become easier to solve than the original one. However, the price for solving an easier problem is that sometimes alternating optimisation may not converge to the same solution as the original problem [59]. For our application, as shown by Figure 5, solving this alternating optimisation problem is able to improve performance compared to the conventional label fusion method, where label fusion and image registration are considered separately.

To conclude, we have proposed a patch-based label fusion model. One of our contributions is that the patch-based model is formulated in a probabilistic Bayesian framework. It is an extension of the probabilistic model in [23] in the way that multiple patches, instead of a single voxel, are extracted from each atlas to account for the registration error. This extension allows us to find the connection between the probabilistic label fusion model and the recently proposed patch-based segmentation method. The second contribution is that label information is incorporated into image registration to improve registration accuracy. Experimental results show that registration refinement improves segmentation accuracy, in terms of both the Dice overlap metric and surface-to-surface distance. The method produces reliable clinical indices which are in good agreement with the manual measurements. It can provide useful information for clinicians in cardiac disease diagnosis.

Acknowledgement

This work was supported by the ElectroCardioMaths Programme of the Imperial British Heart Foundation Centre of Research Excellence, BHF grant RG/10/11/28457, and NIHR Biomedical Research Centre funding. The authors gratefully acknowledge Dr Tim Dawes, Dr Reza S. Razavi for providing the CMR data set, Dr Caroline Petitjean for providing the RV dataset and her warm-hearted help in evaluating segmentation performance, Dr Yingsong Zhang for helpful discussion. Finally, the authors would like to thank the anonymous reviewers for their constructive comments in improving the manuscript.

References

[1] J. Cousty, L. Najman, M. Couprie, S. Clément-Guinaudeau, T. Goissen, and J. Garot. Segmentation of 4D cardiac MRI: Automated method based on spatio-temporal watershed cuts. *Image and Vision Computing*, 28(8):1229–1243, 2010.

[2] M. Jolly. Fully automatic left ventricle segmentation in cardiac cine MR images using registration and minimum surfaces. In *The MIDAS Journal - Cardiac MR Left Ventricle Segmentation Challenge*, 2009.

[3] M. Lorenzo-Valdés, G.I. Sanchez-Ortiz, A.G. Elkington, R.H. Mohiaddin, and D. Rueckert. Segmentation of 4D cardiac MR images using a probabilistic atlas and the EM algorithm. *Medical Image Analysis*, 8(3):255–265, 2004.

[4] M. Lynch, O. Ghita, and P.F. Whelan. Segmentation of the left ventricle of the heart in 3-D+t MRI data using an optimized nonrigid temporal model. *IEEE Transactions on Medical Imaging*, 27(2):195–203, 2008.

[5] C. Petitjean and J.N. Dacher. A review of segmentation methods in short axis cardiac MR images. *Medical Image Analysis*, 15:169–184, 2011.

[6] H. Zhang, A. Wahle, R.K. Johnson, T.D. Scholz, and M. Sonka. 4-D cardiac MR image analysis: Left and right ventricular morphology and function. *IEEE Transactions on Medical Imaging*, 29(2):350–364, 2010.

[7] Y. Zhu, X. Papademetris, A.J. Sinusas, and J.S. Duncan. Segmentation of the left ventricle from cardiac MR images using a subject-specific dynamical model. *IEEE Transactions on Medical Imaging*, 29(3):669–687, 2010.

[8] X. Zhuang, K.S. Rhode, R.S. Razavi, D.J. Hawkes, and S. Ourselin. A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI. *IEEE Transactions on Medical Imaging*, 29(9):1612–1625, 2010.

[9] P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick, and G. Wright. Evaluation framework for algorithms segmenting short axis cardiac MRI. *MICCAI Workshop: Cardiac MR LV Segmentation Challenge*, 2009.

[10] A. Suinesiaputra, B.R. Cowan, Finn J.P., C.G. Fonseca, A.H. Kadish, D.C. Lee, P. Medrano-Gracia, S.K. Warfield, W. Tao, and A.A. Young. Left ventricular segmentation challenge from cardiac MRI: a collation study. *MICCAI Cardiac MR LV Segmentation Challenge*, 2011.

[11] Caroline Petitjean, Su Ruan, Damien Grosgeorge, Jérôme Caudron, and Jean-Nicolas Dacher. Right ventricle segmentation in cardiac MRI: a MICCAI12 challenge. In *MICCAI 2012 Right Ventricle Segmentation Challenge Workshop*, 2012.

[12] Jérôme Caudron, Jeannette Fares, Valentin Lefebvre, Pierre-Hugues Vivier, Caroline Petitjean, and Jean-Nicolas Dacher. Cardiac MRI assessment of right ventricular function in acquired heart disease: factors of variability. *Academic Radiology*, 19(8):991–1002, 2012.

[13] 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(1):115–126, 2006.

[14] T. Rohlfing, D.B. Russakoff, and C.R. Maurer. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. *IEEE Transactions on Medical Imaging*, 23(8):983–994, 2004.

[15] 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(4):1428–1442, 2004.

[16] C. Svarer, K. Madsen, S.G. Hasselbalch, L.H. Pinborg, S. Haugbøl, V.G. Frøkjær, S. Holm, O.B. Paulson, and G.M. Knudsen. MR-based automatic delineation of volumes of interest in human brain PET images using probability maps. *NeuroImage*, 24(4):969–979, 2005.

[17] P. Aljabar, RA Heckemann, A. Hammers, JV Hajnal, and D. Rueckert. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage*, 46(3):726–738, 2009.

[18] J.M.P. Lotjonen, R. Wolz, J.R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, and D. Rueckert. Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage*, 49(3):2352–2365, 2010.

[19] E.M. van Rikxoort, I. Išgum, Y. Arzhaeva, M. Staring, S. Klein, M.A. Viergever, J.P.W. Pluim, and B. van Ginneken. Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus. *Medical Image Analysis*, 14(1):39–49, 2010.

[20] X. Artaechevarria, A. Muñoz-Barrutia, and C. Ortiz-de Solorzano. Combination strategies in multi-atlas image segmentation: Application to brain MR data. *IEEE TMI*, 28(8):1266–1277, 2009.

[21] I. Išgum, M. Staring, A. Rutten, M. Prokop, M.A. Viergever, and B. van Ginneken. Multi-atlas-based segmentation with local decision fusion - Application to cardiac and aortic segmentation in CT scans. *IEEE Transactions on Medical Imaging*, 28(7):1000–1010, 2009.

[22] T.R. Langerak, U.A. van der Heide, A.N.T.J. Kotte, M.A. Viergever, M. van Vulpen, and J.P.W. Pluim. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). *IEEE TMI*, 29(12):2000–2008, 2010.

[23] M.R. Sabuncu, B.T.T. Yeo, K. Van Leemput, B. Fischl, and P. Golland. A generative model for image segmentation based on label fusion. *IEEE Transactions on Medical Imaging*, 29(10):1714–1729, 2010.

[24] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[25] Pierrick Coupé, José Manjón, Vladimir Fonov, Jens Pruessner, Montserrat Robles, and D Collins. Nonlocal patch-based label fusion for hippocampus segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 129–136. Springer, 2010.

[26] P. Coupé, J.V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D.L. Collins. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage*, 54(2):940–954, 2011.

[27] V. Katkovnik, A. Foi, K. Egiazarian, and J. Astola. From local kernel to nonlocal multiple-model image denoising. *International Journal of Computer Vision*, 86(1):1–32, 2010.

[28] A. Buades, B. Coll, and J.M. Morel. Nonlocal image and movie denoising. *Int. Journal of Computer Vision*, 76(2):123–139, 2008.

[29] P. Coupé, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot. An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images. *IEEE TMI*, 27(4):425–441, 2008.

[30] F. Rousseau, P.A. Habas, and C. Studholme. A supervised patch-based approach for human brain labeling. *IEEE Transactions on Medical Imaging*, 30(10):1852–1862, 2011.

[31] H. Wang, J.W. Suh, S. Das, J. Pluta, M. Altinay, and P. Yushkevich. Regression-based label fusion for multi-atlas segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1113–1120, 2011.

[32] T. Tong, R. Wolz, J.V. Hajnal, and D. Rueckert. Segmentation of brain mr images via sparse patch representation. In *MICCAI Workshop: Sparsity Techniques in Medical Imaging*, 2012.

[33] D. Zhang, Q. Guo, G. Wu, and D. Shen. Sparse patch-based label fusion for multi-atlas segmentation. In *MICCAI Workshop: Multimodal Brain Image Analysis*, pages 94–102. Springer, 2012.

[34] H. Wang, S.R. Das, J.W. Suh, M. Altinay, J. Pluta, C. Craige, B. Avants, and P.A. Yushkevich. A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation. *NeuroImage*, 55(3):968–985, 2011.

[35] H. Wang, J. Suh, S. Das, J. Pluta, C. Craige, and P. Yushkevich. Multi-atlas segmentation with joint label fusion. *IEEE Transactions on PAMI*, 35(3):611–623, 2013.

[36] D. Zhang, G. Wu, H. Jia, and D. Shen. Confidence-guided sequential label fusion for multi-atlas based segmentation. *MICCAI 2011*, pages 643–650, 2011.

[37] S. Hu, P. Coupé, J.C. Pruessner, and D.L. Collins. Nonlocal regularization for active appearance model: Application to medial temporal lobe segmentation. *Human Brain Mapping*, 2012.

[38] R. Wolz, C. Chu, K. Misawa, K. Mori, and D. Rueckert. Multi-organ abdominal ct segmentation using hierarchically weighted subject-specific atlases. In *MICCAI 2012*, pages 10–17. Springer, 2012.

[39] V. Fonov, P. Coupé, S.F. Eskildsen, J.V. Manjon, L. Collins, et al. Multi-atlas labeling with population-specific template and non-local patch-based label fusion. In *MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling*, pages 63–66, 2012.

[40] S.F. Eskildsen, P. Coupé, V. Fonov, J.V. Manjón, K.K. Leung, N. Guizard, S.N. Wassef, L.R. Østergaard, and D.L. Collins. BEaST: Brain extraction based on nonlocal segmentation technique. *NeuroImage*, 59(3):2362–2373, 2012.

[41] J.E. Iglesias, M.R. Sabuncu, and K. Van Leemput. A generative model for multi-atlas segmentation across modalities. In *International Symposium on Biomedical Imaging*, pages 888–891. IEEE, 2012.

[42] J. Iglesias, M. Sabuncu, and K. Van Leemput. A generative model for probabilistic label fusion of multimodal data. *MICCAI Workshop: Multimodal Brain Image Analysis*, pages 115–133, 2012.

[43] A. Asman and B. Landman. Non-local STAPLE: an intensity-driven multi-atlas rater model. *MICCAI 2012*, pages 426–434, 2012.

[44] A. Asman and B. Landman. Non-local statistical label fusion for multi-atlas segmentation. *Medical Image Analysis*, 2012.

[45] Hákon Gudbjartsson and Samuel Patz. The Rician distribution of noisy MRI data. *Magnetic Resonance in Medicine*, 34(6):910–914, 1995.

[46] Petter Risholm, Eigil Samset, and William Wells. Bayesian estimation of deformation and elastic parameters in non-rigid registration. *Biomedical Image Registration*, pages 104–115, 2010.

[47] Petter Risholm, Steve Pieper, Eigil Samset, and William Wells. Summarizing and visualizing uncertainty in non-rigid registration. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pages 554–561, 2010.

[48] Ivor JA Simpson, Julia A Schnabel, Adrian R Groves, Jesper LR Andersson, and Mark W Woolrich. Probabilistic inference of regularisation in non-rigid registration. *NeuroImage*, 59(3):2438–2451, 2012.

[49] Ivor JA Simpson, Mark W Woolrich, Jesper LR Andersson, Adrian R Groves, and Julia A Schnabel. Ensemble learning incorporating uncertain registration. *IEEE Transactions on Medical Imaging*, to appear.

[50] J. Ashburner and K.J. Friston. Unified segmentation. *NeuroImage*, 26(3):839–851, 2005.

[51] X. Chen, M. Brady, J.L.C. Lo, and N. Moore. Simultaneous segmentation and registration of contrast-enhanced breast MRI. In *Information Processing in Medical Imaging*, pages 126–137. Springer, 2005.

[52] E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens. An information theoretic approach for non-rigid image registration using voxel class probabilities. *Medical Image Analysis*, 10(3):413–431, 2006.

[53] E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens. A unified framework for atlas based brain image segmentation and registration. In *International Workshop on Biomedical Image Registration*, pages 136–143. Springer, 2006.

[54] 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 Transactions on Medical Imaging*, 18(8):712–721, 1999.

[55] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004, 2003.

[56] A.F. Frangi, D. Rueckert, J.A. Schnabel, and W.J. Niessen. Automatic construction of multiple-object three-dimensional statistical shape models: Application to cardiac modeling. *IEEE Transactions on Medical Imaging*, 21(9):1151–1166, 2002.

[57] F. Grothues, G.C. Smith, J.C.C. Moon, N.G. Bellenger, P. Collins, H.U. Klein, and D.J. Pennell. Comparison of interstudy reproducibility of cardiovascular magnetic resonance with two-dimensional echocardiography in normal subjects and in patients with heart failure or left ventricular hypertrophy. *The American Journal of Cardiology*, 90(1):29–34, 2002.

[58] L.G. Nyul, J.K. Udupa, and X. Zhang. New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150, 2000.

[59] J. Bezdek and R. Hathaway. Some notes on alternating optimization. *Advances in Soft Computing-AFSS 2002*, pages 288–300, 2002.