Multi-atlas Segmentation with Learning-Based Label Fusion

Hongzhi Wang, Yu Cao, and Tanveer Syeda-Mahmood

IBM Almaden Research Center

Abstract. Although multi-atlas segmentation techniques have been producing impressive results for many medical image segmentation problems, most label fusion methods developed so far rely on simple statistical inference models that may not be optimal for inference in high-dimensional feature space. To address this problem, we propose a novel scheme that allows more effective usage of advanced machine learning techniques for patch-based label fusion. Our key novelty is using image registration to guide training sample selection for more effective learning. We demonstrate the power of this new technique in cardiac segmentation using clinical 2D ultrasound images and show superior performance over multi-atlas segmentation and machine learning-based segmentation.

1 Introduction

Multi-atlas segmentation has demonstrated outstanding performance for a wide range of medical image segmentation problems. One key ingredient to its success is that deformable registration can accurately align anatomical structures across subjects for reliable label propagation. With accurate structure alignment, simple label fusion methods such as similarity-based local weighted voting [6,11,15] often can produce state of the art performance.

Although more powerful learning and classification techniques other than weighted voting have been developed in machine learning research, applying advanced machine learning techniques to aid label fusion has not been extensively studied. In some recent work, [13] applies adaboost classification as a postprocessing step to reduce errors produced by multi-atlas segmentation. Similarly, [5] employed random forest classification to reduce ambiguities produced by multi-atlas segmentation. Random forest is also employed in [18] for atlas encoding.

A common limitation of the above mentioned methods is that the classifiers are all trained without taking advantage of registration-based structure alignment, i.e. the key advantage of multi-atlas segmentation. To address this limitation, we explore a new scheme for combining learning techniques with multi-atlas segmentation. Our key novelty lies in an image registration based training sample selection strategy for more effective learning. Similar to the spatially varying image similarity based local weight voting approach, we propose a spatially varying training sample selection strategy that aims to only apply training samples that are anatomically most relevant to the target testing sample for segmenting the target sample. This is achieved by selecting training samples within a

G. Wu et al. (Eds.): MLMI 2014, LNCS 8679, pp. 256–263, 2014.

local neighborhood surrounding the target testing sample after registering and warping atlases into the target image.

We implement our method with random forest and conduct validation in a challenging cardiac segmentation application using clinical four-chamber view 2D echocardiography. We demonstrate promising improvement over the state of the art label fusion method and random forest based segmentation.

2 Method

2.1 Background in Patch-Based Multi-atlas Label Fusion

In this section, we briefly describe multi-atlas segmentation. Let T_F be a target image to be segmented and $A^1 = (A^1_F, A^1_S), ..., A^n = (A^n_F, A^n_S)$ be n atlases, warped to the space of the target image by deformable registration. A^i_F and A^i_S denote the i_{th} warped atlas image and manual segmentation. Each A^i_S is a candidate segmentation for the target image. Label fusion combines these candidate segmentations to produce the final solution.

One simple and highly effective label fusion method is based on weighted voting. For instance, the combined votes for label l are:

$$\hat{p}(l|x, T_F) = \sum_{i=1}^{n} w^i_x p(l|x, A^i) \tag{1}$$

where x indexes through image locations. $\hat{p}(l|x, T_F)$ is the estimated label posterior for the target image. $p(l|x, A^i)$ is the probability that A^i votes for label l at x, with $\sum_{l\in\{1,...,L\}} p(l|x, A^i) = 1$. L is the total number of labels. w^i_x is a local weight assigned to the i_{th} atlas, with $\sum_{i=1}^{n} w^i_x = 1$. The voting weights are typically determined based on the quality of registration produced for each atlas such that more accurately registered atlases are weighted more heavily in producing the final solution.

Patch-based label fusion. For estimating registration/segmentation accuracy, patch-based approaches are among the most effective techniques. For this task, most methods apply similarity metrics typically employed by image-based registration, such as sum of squared distance (SSD) and normalized cross correlation (NCC) computed over local image patches. For instance, when SSD and a Gaussian weighting model are used [11], the voting weights in (1) can be estimated by $w^i_x = \frac{1}{Z(x)} exp\left(-\sum_{y\in\mathcal{N}(x)} \left[A^i_F(y) - T_F(y)\right]^2 / \sigma\right)$, where σ is a model parameter. $\mathcal{N}(x)$ defines the image patch, which is a neighborhood surrounding x, and $Z(x)$ is a normalization constant.

Although the above approach can provide reasonable estimation about registration accuracy for each warped atlas, its contribution for remedying the registration error is limited. To more effectively remedy registration errors, atlas patches within a local searching neighborhood of the registered correspondence could all be considered as the potential corresponding patch for a target patch and are applied for label fusion in patch-based label fusion methods [4,10,15].

2.2 Limitations of Current Patch-Based Label Fusion Methods

Patch-based label fusion could be interpreted as a regression or interpolation problem [9,14], where the goal is to predict the segmentation label for each target voxel given its surrounding image patch. All potential corresponding image patches from the warped atlases provide observed data for this regression task. Given that this regression problem is performed in a high-dimensional feature space, the simple distance metric employed in current patch-based label fusion methods, e.g. the Euclidean metric used above, could be inadequate for accurately characterizing the feature space. This problem is less critical when image registration can be reliably computed. As shown in an empirical study [16], simple metrics such as the Euclidean metric does a good job differentiating small registration errors, however the accuracy of predicting large registration errors quickly drops as the registration error increases. Hence, employing simple metrics in patch-based label fusion becomes more problematic when image registrations are poorly computed.

2.3 Random Forest Based Label Fusion

To address this limitation, we propose to employ more powerful learning techniques for patch-based label fusion. In this paper, we investigate the usage of random forest for this task.

A random forest is an ensemble of decision trees [3]. Each non-leaf node in a decision tree performs a test, e.g. the comparison of a feature value to a given threshold. During training, training data are used to build each decision tree. During testing, a testing data is sent to the root node of each decision tree. Based on the test at the node, the data is sent to either its left or right child node. This process is repeated until a leaf node is reached in a tree. The class distribution of all training samples located in the leaf node is interpreted as the probability that the testing data should be assigned to each class. The final class probability is obtained by averaging the class distributions from all decision trees. Random forest has demonstrated impressive performance in image segmentation [18].

Inspired by the highly successful spatially varying weighted voting scheme, we propose to train spatially varying local random forest classifiers for label fusion. As in patch-based label fusion [4,10], given a target patch, all atlas patches located in a small neighborhood of the registered correspondence are applied for training a local random forest classifier, which is then applied for predicting labels for the target patch. To facilitate our comparison with previous patch-based label fusion, we apply pixel intensity values within each image patch as features to predict the patch's central pixel's label. In our experiment, we apply a $(2r_s + 1) \times (2r_s + 1)$ square-shaped sampling neighborhood specified by the radius r_s for our 2D cardiac ultrasound images. We also use a square-shaped patch specified by a radius r for feature extraction.

Ideally, a distinct random forest classifier should be trained for segmenting each target voxel using warped atlas samples surrounding the target voxel. However, this requirement increases the computational cost. To make our study more practical, we apply the trained classifier to predict segmentation labels for each voxel located in the sampling window in the target image. In addition, we train classifiers on overlapping sampling windows. Centers of the sampling windows are located on a 2D grid with the distance between two neighboring grid nodes in each row and each column equal $\frac{r_s}{2}$. Classification results from overlapping classifiers are averaged to generate the final solution.

2.4 Relation to Training Sample Selection in Machine Learning

In machine learning research, it is well known that not all training samples are always equally important for any given learning task. Choosing the most relevant training samples for any specific classification task is a highly effective technique for improving the learning performance. In general, rules for choosing the most relevant training samples are application dependent. In medical image segmentation, one intuitive rule for choosing relevant training samples is based on their anatomical locations.

Although machine learning methods usually employ more powerful statistical inference techniques than what are employed by current patch-based label fusion methods, existing learning-based segmentation methods still largely ignore the valuable anatomical information encoded in medical images. In contrast, multi-atlas segmentation stands in the opposite extreme in terms of how anatomical information is incorporated for reaching solutions. For instance, it is common that machine learning based methods apply mixed training data sampled from different anatomical area for making segmentation decisions, while multi-atlas segmentation significantly simplifies the problem by applying spatially-varying training data that are anatomically more relevant to the testing data obtained from image registration. Due to this distinction, the best brain segmentation performance achieved by machine learning techniques, e.g. [12,18], are still well below those produced by multi-atlas segmentation [7,1]. In this aspect, *the key advantage of our method lies in combining the complementary advantages of machine learning and multi-atlas segmentation.*

3 Experiments

We conduct experimental study on cardiac segmentation using apical four-chamber view 2D echocardiography. 2D echocardiography is a common modality for diagnosis in clinical practice. Anatomical structure labeling will assist cardiac disease diagnosis by providing geometrical and morphological statistics. This is an ideal application for demonstrating the advantage of our machine learning based label fusion technique. Image registration on echocardiography are challenging as the images are noisy and the anatomical structure deformation among different subjects are large. Hence, label fusion needs to accommodate large registration errors. Furthermore, different anatomical regions often share similar intensity profiles in echocardiography, making image feature based machine learning techniques less effective.

3.1 Data and Experiment Setup

Our dataset consists of a total of 50 patients with a variety of cardiac diseases such as aneurysms, dilated cardiomyopathy and hypertrophies. Each image is manually labeled with the following nine structures: Chamber Junction (CJ), Inter Ventricle Septum (IVS), Left Ventricle (LV), Mitral Valve (MV), Left Atrium (LA), Inter Atrium Septum (IAS), Right Atrium (RA), Tricuspid Valve (TV) and Right Ventricle (RV). We conducted a 5-fold cross-validation. Hence, the dataset is randomly divided into 5 equal size non-overlap groups. Each group is treated as the testing set and the remaining groups are treated as training set in each of the five cross-validation experiments. The results below are summarized over the five cross-validation experiments. In our experiments, we applied sampling windows with $r_s = 5$ for our method.

Deformable image registration. The global image-based registration between each pair of images were performed through sequentially optimizing translation, rigid body, affine and deformable transforms between the registered images. Deformable registration was performed using the greedy diffeomorphic Symmetric Normalization (SyN) algorithm [2] implemented by the Advanced Normalization Tools (ANTs) software package. The Mattes mutual information metric was applied for the registration task. Multi-scale optimization was applied. Three resolution levels with maximum 200 iterations at the coarse and middle levels and 100 iterations at the fine level were applied.

Random Forest setup. We applied the random forest package implemented in R [8] with the default parameter setting, i.e. 500 trees. Using this implementation, our method usually segments each image in about 10 minutes.

Benchmark methods. For comparison, we evaluated joint label fusion (JLF) [15]. This method is one of the state of the art methods for patch-based local weighted voting label fusion and is a consistent top performer in both MICCAI grand challenges on multi-atlas segmentation held in 2012 and 2013 [7,1]. For this study, we applied the authors' implementation that is distributed through the ANTs software package with default parameters, i.e. 5×5 image patches for local image similarity estimation, 7×7 local searching windows and model parameter $\sigma = 2$. As another baseline performance, we also computed the segmentation results produced by majority voting (MV) and by the STAPLE algorithm [17].

In the second comparison, we compare with the segmentation performance produced by the classical usage of random forest for image segmentation (RF). We train a single random forest classifier using the training samples from all atlases without warping them into the target image space and apply this classifier to segment testing images. In addition to the intensity feature extracted from each pixel's surrounding patch, we also include relative spatial location of each training sample with respect to the mass center of the scanned view as an additional feature. To facilitate a direct comparison with other tested methods, we did not include any other features for random forest classification.

Table 1. Segmentation performance of our random forest label fusion (RFLF) method compared with other methods. Results are measured using the Dice similarity coefficient $(2|A \cap B|/|A| + |B|)$.

anatomical regions	MV	STAPLE	RF	JLF	RFLF
CJ	0.53±0.28	0.43±0.21	0.49±0.24	0.57±0.25	0.64±0.18
IVS	0.67±0.27	0.62±0.18	0.72±0.13	0.73±0.19	0.78±0.09
LV	0.77±0.14	0.73±0.17	0.82±0.07	0.80±0.11	0.82±0.08
MV	0.30±0.25	0.34±0.28	0.20±0.10	0.44±0.25	0.50±0.18
LA	0.74±0.20	0.69±0.19	0.73±0.14	0.78±0.17	0.79±0.14
IAS	0.51±0.29	0.38±0.27	0.18±0.14	0.57±0.23	0.59±0.21
RA	0.70±0.21	0.60±0.26	0.69±0.17	0.73±0.20	0.75±0.16
TV	0.07±0.12	0.17±0.21	0.02±0.03	0.17±0.18	0.21±0.19
RV	0.68±0.18	0.53±0.24	0.68±0.17	0.72±0.15	0.72±0.14
Overall	0.55	0.50	0.50	0.61	0.65

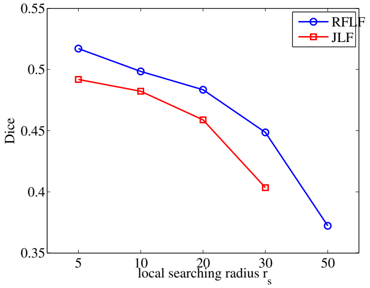

Fig. 1. Segmentation accuracy (in terms of average Jaccard index) of joint label fusion and our random forest based label fusion method with respective to the size of local searching windows

Results. Table 1 summarizes the segmentation performance produced by each method. The performance of the classical machine learning approach that learns a single random forest classifier to assign labels for the entire testing image is clearly below those of multi-atlas segmentation methods. In contrast, applying random forest for local patch-based label prediction produced a significant improvement over the state of the art label fusion method (with $p < 0.05$ on the paired Students t-test compared with JLF and $p < 0.001$ compared with the remaining evaluated methods). This result clearly demonstrates: 1) the simple metric based patch label fusion method is inadequate for our application; 2) the registration based spatially varying sample selection scheme significantly improved the performance of random forest.

Fig. 1 shows the performance of joint label fusion and our random forest based label fusion method with respect to the size of local searching windows. Both methods' performance dropped as the local searching radius increases. This result is expected because larger local searching/sampling windows complicate the label fusion/classification problem by adding more irrelevant samples into

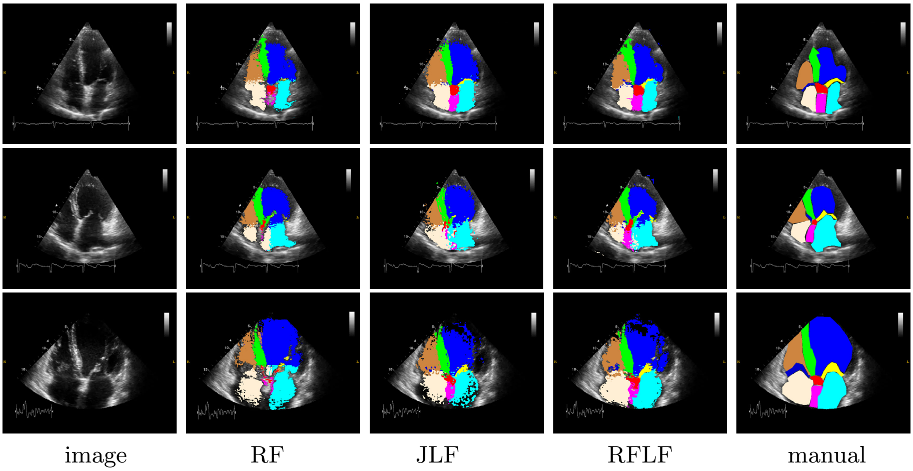

Fig. 2. Segmentation results by different methods. Red: CJ; Green: IVS; Blue: LV; Yellow: MV; Sky blue: LA; Pink: IAS; Light brown: RA; Deep blue: TV; Brown: RV.

consideration. This result also indicates that training separate classifiers for distinct anatomical structures, such as in [13,5], is suboptimal because one anatomical structure may still be large enough to include samples that are not strongly relevant with each other for classification purpose. See Fig. 2 for some segmentation examples.

4 Discussion and Conclusions

We introduced a novel scheme for combining the complementary advantages of multi-atlas segmentation with more general machine learning techniques. The key idea is to use image registration to generate spatially varying training sample selection for more effective learning. In our experiments of cardiac segmentation using four chamber view 2D echocardiography, we demonstrated that the registration-based spatially varying sample selection method significantly improves classification accuracy for random forest. By including more descriptive features or by applying postprocessing methods such as [13,5], we expect further prominent improvement in the segmentation performance. In future work, we will also conduct validation on broader applications with different registration accuracy levels.

One common implementation to make multi-atlas segmentation more practical is to preregister all atlases to a common template space. Given a new target image, only one registration from the target image to the template is required. Although it significantly reduces the registration burden, it also compromises the overall registration accuracy. Since our experiments show that our machine learning based label fusion method is more robust to registration errors, it is especially suitable to be implemented through the common template strategy.

References

1. Asman, A., Akhondi-Asl, A., Wang, H., Tustison, N., Avants, B., Warfield, S.K., Landman, B.: Miccai 2013 segmentation algorithms, theory and applications (SATA) challenge results summary. In: MICCAI 2013 Challenge Workshop on Segmentation: Algorithms, Theory and Applications. Springer (2013)
2. Avants, B., Epstein, C., Grossman, M., Gee, J.: Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12(1), 26–41 (2008)
3. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
4. Coupe, P., Manjon, J., Fonov, V., Pruessner, J., Robles, N., Collins, D.: Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54(2), 940–954 (2011)
5. Han, X.: Learning-boosted label fusion for multi-atlas auto-segmentation. In: Wu, G., Zhang, D., Shen, D., Yan, P., Suzuki, K., Wang, F. (eds.) MLMI 2013. LNCS, vol. 8184, pp. 17–24. Springer, Heidelberg (2013)
6. Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M., van Ginneken, B.: Multi-atlas-based segmentation with local decision fusion–application to cardiac and aortic segmentation in CT scans. IEEE Trans. on MI 28(7), 1000–1010 (2009)
7. Landman, B., Warfield, S. (eds.): MICCAI 2012 Workshop on Multi-Atlas Labeling. CreateSpace (2012)
8. Liaw, A., Wiener, M.: Classification and regression by randomforest. R News 2(3), 18–22 (2002), `http://CRAN.R-project.org/doc/Rnews/`
9. Rohlfing, T., Brandt, R., Menzel, R., Russakoff, D.B., Maurer Jr., C.R.: Quo vadis, atlas-based segmentation? In: The Handbook of Medical Image Analysis–Volume III: Registration Models, pp. 435–486 (2005)
10. Rousseau, F., Habas, P.A., Studholme, C.: A supervised patch-based approach for human brain labeling. IEEE TMI 30(10), 1852–1862 (2011)
11. Sabuncu, M., Yeo, B., Leemput, K.V., Fischl, B., Golland, P.: A generative model for image segmentation based on label fusion. IEEE TMI 29(10), 1714–1720 (2010)
12. Tu, Z., Zheng, S., Yuille, A., Reiss, A., Dutton, R., Lee, A., Galaburda, A., Dinov, I., Thompson, P., Toga, A.: Automated extraction of the cortical sulci based on a supervised learning approach. IEEE TMI 26(4), 541–552 (2007)
13. Wang, H., Das, S., Suh, J.W., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P.: 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)
14. Wang, H., Suh, J.W., Das, S., Pluta, J., Altinay, M., Yushkevich, P.: Regression-based label fusion for multi-atlas segmentation. In: CVPR (2011)
15. Wang, H., Suh, J.W., Das, S., Pluta, J., Craige, C., Yushkevich, P.: Multi-atlas segmentation with joint label fusion. IEEE Trans. on Pattern Analysis and Machine Intelligence 35(3), 611–623 (2013)
16. Wang, H., Yushkevich, P.A.: Spatial bias in multi-atlas based segmentation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 909–916. IEEE (2012)
17. Warfield, S., Zou, K., Wells, W.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE TMI 23(7), 903–921 (2004)
18. Zikic, D., Glocker, B., Criminisi, A.: Atlas encoding by randomized forests for efficient label propagation. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part III. LNCS, vol. 8151, pp. 66–73. Springer, Heidelberg (2013)