Learning-Boosted Label Fusion for Multi-atlas Auto-Segmentation

Xiao Han

Elekta Inc., St. Louis, MO, USA
xiao.han@elekta.com

Abstract. Structure segmentation of patient CT images is an essential step for radiotherapy planning but very tedious if done manually. Atlas-based auto-segmentation (ABAS) methods have shown great promise for getting accurate segmentation results especially when multiple atlases are used. In this work, we aim to further improve the performance of ABAS by integrating it with learning-based segmentation techniques. In particular, the Random Forests (RF) supervised learning algorithm is applied to construct voxel-wise structure classifiers using both local and contextual image features. Training of the RF classifiers is specially tailored towards structure border regions where errors in ABAS segmentation typically occur. The trained classifiers are applied to re-estimate structure labels at "ambiguous" voxels where labels from different atlases do not fully agree. The classification result is combined with traditional label fusion to achieve improved accuracy. Experimental results on H&N images and ribcage segmentation show clear advantage of the proposed method, which offers consistent and significant improvements over the baseline method.

Keywords: atlas-based segmentation, machine learning, label fusion, random forests, radiotherapy planning, CT image.

1 Introduction

Structure segmentation of patient CT images is an essential step for radiotherapy planning. Although manual contouring by human experts is still the common standard for high quality segmentation in clinics, it is tedious, time-consuming and suffers from large intra- and inter- rater variability.

Automated segmentation of CT images is a very challenging problem due to image noise and other artifacts, as well as limited image contrast for most soft-tissue structures. In recent years, atlas-based auto-segmentation (ABAS) methods have shown great promise in helping solve the problem and been applied in commercial products [1-2]. Although the segmentation results still need be edited manually before they can be used clinically, ABAS methods have been proven to be able to greatly reduce manual labor and improve contouring consistency [1].

The basic principle of ABAS is to perform segmentation of a novel patient image using expert-labeled images, called atlases. After aligning the new image to the atlas image through image registration, atlas structure labels can be mapped to the patient

G. Wu et al. (Eds.): MLMI 2013, LNCS 8184, pp. 17–24, 2013.

image to get the automatic segmentation result. Large anatomical variation among different subjects often limits the accuracy of ABAS if only a single atlas is used. Thus, it becomes common standard to use multiple atlases, where each atlas is first applied independently and their results are combined in the end through label fusion.

Even with multi-atlas and label fusion, accuracy of ABAS is still heavily dependent on performance of image registration. Rather than relying on image registration alone, in this work we aim to combine the strength of multi-atlas ABAS with that of learning-based image segmentation techniques in order to get much improved accuracy. In the method developed here, we apply Random Forests (RF) – a state-of-the-art supervised learning algorithm (cf. [3]) to construct a voxel classifier for each structure using the existing atlases as training data. The RF algorithm can effectively handle a large number of training data with high data dimension, which allows us to explore a large number of image features to fully capture both local and contextual image information. We also specially tailor the training of the RF algorithm to focus on structure border regions where errors in ABAS typically occur. After a standard multi-atlas label fusion is performed, the RF classifier(s) are applied to re-estimate the label probability for voxels where labels mapped from different atlases do not fully agree. The RF result is then combined with the initial label fusion to get the final structure segmentation.

There are some related works in the literature. The RF method itself has been applied for structure localization and lesion segmentation problems (cf. [3]). In [4], Powell et al used ANN-based voxel classifier to improve brain structure segmentation from a probabilistic atlas. Nie and Shen [6] designed a SVM-guided deformable surface model to refine structure surface segmentation of mouse brain images. Hao et al [5] applied a Lagrangian SVM algorithm to train massive localized voxel classifiers on the fly for hippocampus segmentation with multiple atlases. The SVM classification was directly used as final result instead of being combined with traditional label fusion. The method can be slow since one classifier is built for each voxel, and only a small number of features were used. Another recent work [7] combined multi-atlas ABAS with a simpler kNN classifier for brain image segmentation. Only six local intensity values were used as voxel features, which is unlikely to produce accurate voxel classification for CT images. Some other works [9, 11] applied machine learning driven statistical shape models for structure detection and segmentation in either CT or MR images. As a competing method, ABAS has its own advantages. For example, spatial structure relationship and full image information are implicitly taken into account during atlas registration. But shape model can also be incorporated to further improve ABAS accuracy.

2 Methods

The proposed method integrates learning-based voxel classification at the label fusion stage within a multi-atlas ABAS framework. We first present the underlying multi-atlas ABAS method and then discuss our design of RF-based voxel classification and the incorporation of it to improve the label fusion accuracy.

2.1 Multi-atlas ABAS Method

Fig. 1 summarizes the workflow of a basic multi-atlas ABAS procedure that we adopt in this work, where the segmentation of a new subject is computed by applying multiple atlases separately and then combining the individual segmentation results through label fusion.

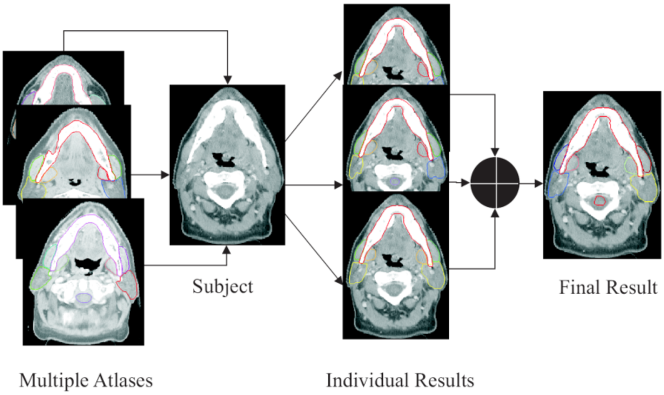

Fig. 1. Overall workflow of ABAS with multiple atlases and label fusion

The registration of each individual atlas to the subject image is performed using a hierarchical atlas registration method we previously developed in [8]. The method computes first a global mutual information (MI) linear registration followed by two non-linear registration steps with gradually increasing degrees-of-freedom. Structure surface information from the atlas is incorporated into the deformation field regularization to improve both the robustness and the accuracy of atlas registration.

The popular STAPLE method [10] is used to combine multiple structure label maps from the different atlases. Although the STAPLE method appears to have a weakness in that it does not make use of image intensity information, we found it work well for CT images. We have also tried various intensity-weighted label fusion methods but found that the improvement is minimal or none since local intensity information is often ambiguous for CT images and sensitive to common CT artifacts.

2.2 RF Voxel Classification

Learning-based classification methods offer an alternative approach for object detection and segmentation. Voxel classifiers, once trained, can predict the structure label of a new image based on discriminative features computed at each voxel location. Voxel classification alone also has its limitations, and is thus often used together with other techniques such as statistical shape models [9, 11]. In this work, we apply learning-based voxel classification to complement ABAS – a registration-based approach. The expert-labeled atlases available in multi-atlas ABAS naturally serve as training data for building the voxel classifier.

We train a RF classifier for each structure to predict the probability of a voxel as belonging to the specific structure. Although RF can easily handle multi-class classification as well, there is minimal benefit for the structures we consider in this work since they are not directly adjacent to each other. RF is a state-of-the-art supervised learning method and often considered to have better generalization power than SVM or boosting [3]. It achieves high generalization by growing an ensemble of independent decision trees on random subsets of the training data and by randomizing the features made available to each tree node during training. The RF algorithm is very efficient and can effectively deal with a very large number of features. In addition, RF also estimates the confidence of the prediction as a by-product of the training process. We apply the standard RF algorithm in this work, where decision stumps are used as weak classifiers and the Gini index is used as the impurity criterion.

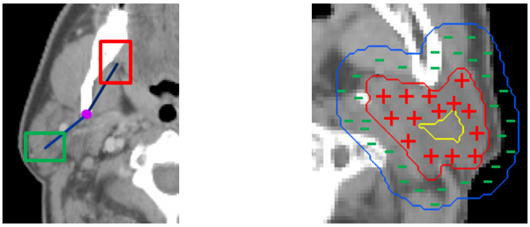
(a) (b)

Fig. 2. Illustration of RF training. (a) contextual feature definition; (b) training samples selection.

Using image intensity alone is insufficient for accurate voxel classification. Relying on point-wise intensity comparison to predict voxel correspondence is also a major limiting factor of image registration algorithms. Instead, we employ a large set of features in order to fully capture both local and contextual information at each image point, which include:

- Image intensity values – the raw intensity value I and the smoothed ones: $G_\sigma * I$, where G_σ denotes a Gaussian filter with a kernel size (scale) of σ. Three different scales are used in this work: 1.0 mm, 1.7 mm, and 2.5 mm.
- Image gradients $(I_x, I_y, I_z) = \nabla(G_\sigma * I)$ and gradient magnitudes $\|\nabla(G_\sigma * I)\|$ computed at three different scales.
- Eigen-values of the image Hessian matrix $H = \nabla^T \nabla(G_\sigma * I)$, which are again computed at three different scales.
- Image location – the (x, y, z) coordinates of a voxel. The coordinates are normalized first with respect to a common reference frame by aligning each image to a fixed reference image through a coarse B-spline image registration.
- Generalized Haar-like features as proposed in [3], which help capture contextual information. As illustrated in Fig. 2a, each such feature is computed as the mean image property difference over two randomly displaced, asymmetric cubical regions around the voxel: $f = |R_1|^{-1} \sum_{\mathbf{x} \in \boldsymbol{R}_1} F(\mathbf{x}) - |R_2|^{-1} \sum_{\mathbf{x} \in \boldsymbol{R}_2} F(\mathbf{x})$. Such features

can be computed very efficiently with the use of *integral images* (cf. [3]). In this work, $F(\mathbf{x})$ is either the raw image intensity value or the image gradient magnitude. We typically sample 200 random features of each type.

Training data are collected from the atlases with some special consideration. As illustrated in Fig. 2b, training samples are only taken within close proximity to the structure boundary, which helps the RF training focus on voxels close to the structure border – a region where ABAS segmentation error is most likely to occur. We use an 8 mm distance threshold to define the sampling region. Training of the RF classifiers is performed offline and one RF classifier is trained for each structure of interest, as mentioned earlier. The RF algorithm is very fast, and different trees can be trained in parallel. It normally takes less than 20 minutes to build a RF classifier with 50 trees. Note that the RF classifier(s) only need be trained once after the atlases are collected.

2.3 RF-Enhanced Label Fusion

The trained RF classifiers are applied after the standard multi-atlas ABAS computation as described in Section 2.1 is finished. Our goal is to combine the strengths of both techniques. Hence, the RF classifier for each structure is only applied to re-estimate the label probability of "ambiguous" voxels in the original ABAS result, which are voxels where labels mapped from different atlases do not fully agree.

The RF classification result can be combined with the initial ABAS result in different ways. For example, we can use the RF result as an extra input to the STAPLE algorithm. Since both RF and STAPLE produce a probabilistic estimation of voxel labels, we choose to compute the final structure label probability as a simple weighted sum of the RF probability (P_R) and the initial STAPLE estimation (P_S):

$$P = w_R P_R + w_S P_S, \tag{1}$$

where w_R and w_S are the relative weights of the two terms. In this work, we assign a slightly higher weight for the RF result as $w_R = 0.6$, then $w_S = 0.4$. Once the label probability is computed for every voxel of the subject image, it can be thresholded at 0.5 or the 0.5-isosurface be computed to get the final structure segmentation result.

3 Experimental Results

3.1 Head & Neck (H&N) Image Segmentation

In the first experiment, we apply the learning-enhanced multi-atlas ABAS for the segmentation of H&N cancer patient CT images. Ten randomly collected patient images with manual expert segmentation are used as the test data. All images have a voxel size of $0.9375 \times 0.9375 \times 2.5\ \text{mm}^3$. The following 4 structures are considered in this study: the mandible, the brainstem, and the left and the right parotids.

We use a leave-one-out strategy to evaluate the proposed method: for each subject, the remaining subjects are considered as atlases. The Dice similarity coefficient (cf. [1]) is used to quantify the accuracy when comparing automatic segmentation results with original manual labeling.

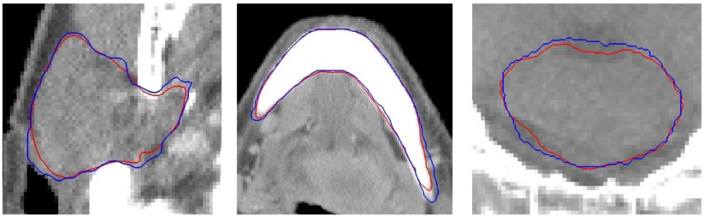

Fig. 3. Qualitative comparison of H&N segmentation results. Blue: STAPLE label fusion results; red: Learning-enhanced label fusion. From left to right: parotid, mandible, and brainstem.

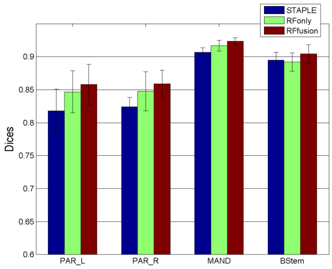

Fig. 4. Quantitative comparison: the bar plots show mean Dice values over 10 subjects for the 4 structures and the error bars indicate one standard deviation

Fig. 3 shows some qualitative comparisons of the segmentation results between the proposed method and the baseline multi-atlas ABAS method with STAPLE label fusion. It can be seen that combining the RF classification clearly improves the segmentation accuracy. Quantitative comparison results are summarized in the box plot of Fig. 4, where the mean and the standard deviation of Dice values for each structure are shown. As can be seen, the learning-enhanced label fusion consistently produces higher accuracy for all 4 structures than the STAPLE method. In addition, the weighted fusion (Eq. (1)) is also more accurate than directly using the RF results (RF-only). Note that the RF-only results still rely on ABAS since RF classification is only computed for the ambiguous voxels as mentioned earlier. It was verified through paired-*t* tests that the improvements of the combined label fusion over both STAPLE and RF-only are statistically significant at the 0.05 level for all 4 structures. Note that some of the remaining segmentation error is inherent to the data due to intra-observer variation, especially for structures with very low contrast such as the brainstem.

The computation time for both the original multi-atlas ABAS method and the RF-enhanced one is quite comparable. It takes about one minute to run a single atlas registration on a desktop computer with an Intel Xeon Quad-core 2.66 GHz CPU and a

NVIDIA GTX 480 graphics card. The STAPLE label fusion takes less than a minute. Computing the RF classification only adds one extra minute, which is about 1/10-th of the total computation time assuming 9 atlases are used.

3.2 Ribcage Segmentation

In the second experiment, we test the proposed method on ribcage segmentation of lung CT images. Expert labeled images from 15 different patients are used as the test data. The image resolution is about $0.9765 \times 0.9765 \times 3\ \text{mm}^3$. We again use the leave-one-out strategy for the validation study, where for each patient image the other 14 are used as atlases for ABAS segmentation and RF training.

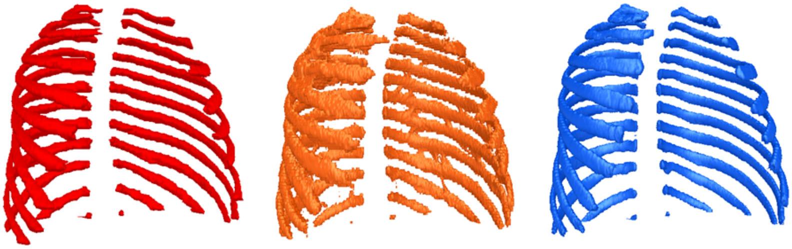

Fig. 5. Illustration of ribcage segmentation results. Left: truth; middle: STAPLE result; right: STAPLE combined with RF classification.

The ribcage segmentation is a difficult problem [12], and turns out to be very challenging for a registration-based segmentation method, i.e., ABAS. It is because the rib bones are all similar to each other and spatially clustered. After linear registration, one rib from the atlas image can partially overlap with two or more ribs from the subject image. Purely intensity-based image registration can never get out of the local optimum of the image similarity function and cause large errors in the final image matching. As a result, the segmentation accuracy is rather low even with 14-atlases, which can be seen from the STAPLE result shown as the middle figure in Fig. 5.

Applying RF-based voxel classification greatly improves the segmentation accuracy, as shown in Fig. 5. Computing the Dice statistics over all 15 patients, we found that the original multi-atlas ABAS with STAPLE label fusion produced Dice values of 0.73 ± 0.03, whereas learning-enhanced label fusion improved the Dice values to 0.86 ± 0.02. The 0.86 overlap ratio is actually very high, considering that the ribs are narrow tube-like structures. This improvement is also statistically significant as verified by the paired-*t* test.

4 Conclusion

We have developed a hybrid multi-atlas ABAS method that effectively combines the strengths of traditional ABAS methods and learning-based segmentation approaches.

Experimental results on H&N CT image segmentation and ribcage segmentation showed significant improvements of the proposed method over the baseline method without the learning-based enhancement. Future work will investigate extra image features such as local binary patterns and region co-variances. We also plan to construct shape priors from the atlases and investigate whether incorporating explicit statistical shape information can further improve the ABAS segmentation accuracy.

Acknowledgment. The author would like to thank Drs. Jan-Jakob Sonke and Barbara Stam from the Netherlands Cancer Institute for providing the lung CT images and the manual ribcage segmentation.

References

1. Pekar, V., Allaire, S., Kim, J., Jaffray, D.A.: Head and Neck Auto-segmentation Challenge. In: van Ginneken, B., Murphy, K., Heimann, T., Pekar, V., Deng, X. (eds.) Medical Image Analysis for the Clinic: A Grand Challenge, pp. 273–280. Springer, Heidelberg (2010)
2. Rohlfing, T., Brandt, R., Menzel, R., Russakoff, D.B., Maurer Jr., C.R.: Quo Vadis, Atlas-based segmentation? In: Suri, J., Wilson, D., Laxminarayan, S. (eds.) The Handbook of Medical Image Analysis. Kluwer (2005)
3. Criminisi, A., Shotton, J.: Decision forests for computer vision and medical image analysis. Springer, London (2013)
4. Powell, S., Magnotta, V.A., Johnson, H., Jammalamadaka, V.K., Andreasen, N.C., Pierson, R.: Registration and machine learning based automated segmentation of subcortical and cerebellar brain structures. NeuroImage 39, 238–247 (2008)
5. Hao, Y., Liu, J., Duan, Y., Zhang, X., Yu, C., Jiang, T., Fan, Y.: Local label learning (L3) for multi-atlas based segmentation. In: Proc. SPIE, vol. 8314, p. 83142E (2012)
6. Nie, J., Shen, D.: Automated segmentation of mouse brain images using multi-atlas multi-ROI deformation and label fusion. Neuroinformatics 11, 35–45 (2013)
7. Srhoj-Egekher, V., Benders, M.J.N.L., Kersbergen, K.J., Viergever, M.A., Isgum, I.: Automatic segmentation of neonatal brain MRI using atlas based segmentation and machine learning approach. In: MICCAI Grand Challenge: Neonatal Brain Segmentation (2012)
8. Han, X., Hoogeman, M., Levendag, P., Hibbard, L., Teguh, D., Voet, P., Cowen, A., Wolf, T.: Atlas-based auto-segmentation of head and neck CT images. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part II. LNCS, vol. 5242, pp. 434–441. Springer, Heidelberg (2008)
9. Zheng, Y., Barbu, A., Georgescu, B., Scheuering, M., Comaniciu, D.: Four-chamber heart modeling and automatic segmentation for 3D cardiac CT volumes using marginal space learning and steerable features. IEEE Trans. Med. Imag. 27, 1668–1681 (2008)
10. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans. Med. Imag. 23, 903–921 (2004)
11. Tu, Z., Narr, K.L., Dollar, P., Dinov, I., Thompson, P.M., Toga, A.W.: Brain anatomical structure segmentation by hybrid discriminative/generative models. IEEE Trans. Med. Imag. 27, 495–508 (2008)
12. Wu, D., Liu, D., Puskas, Z., Lu, C., Wimmer, A., Teitjen, C., Soza, G., Zhou, S.K.: A learning based deformable template matching method for automatic rib centerline extraction and labeling in CT images. In: Proc. CVPR 2012 (2012)