Regression forest region recognition enhances multi-atlas spleen labeling

Bo Li¹, Swetasudha Panda², Zhoubing Xu², Andrew J. Asman², Peter L. Shanahan³, Richard G. Abramson³, Bennett A. Landman¹²³

¹ Computer Science, ² Electrical Engineering, ³ Radiology and Radiological Sciences, Vanderbilt University, Nashville, TN, USA 37235

{bo.li.2, swetasudha.panda, zhoubing.xu, andrew.j.asman, peter.l.shanahan, richard.abramson, bennett.landman}@vanderbilt.edu

Abstract. Complex and variable abdominal anatomy poses considerable challenges for automated segmentation of the spleen. Herein, we propose to use “regression forests region recognition” (RFRR) to identify bounding boxes around a region of interest (e.g., the spleen) prior to image-based registration. Within these pre-identified regions, we apply multi-atlas segmentation using deformable registration. By focusing on reduced regions of interest, fewer structures confound registration and errors are minimized. The primary contributions of this work are: (1) to extend the existing regression forest framework to include local features; (2) to combine automated regional localization with multi-atlas segmentation, and (3) to accurately segment spleen volumes on clinically acquired computed tomography (CT) data. This approach achieves Dice similarity coefficient (DSC) of 0.82±0.13 and mean surface distance of 7.49±10.62 mm versus a traditional multi-atlas method (without reduced regions of interest) with mean DSC of 0.66±0.21 and mean surface distance of 8.67±13.35 mm.

Keywords: Regression forests, region recognition, multi-atlas, spleen segmentation, bounding box

1 Introduction

Multi-region detection and localization of 3-D anatomical structures is a critical challenge in medical image processing. Diverse region detection and recognition approaches have been proposed. Template matching has yielded segmentation based on the geometrically meaningful features [1, 2], but these approaches often have difficulties capturing invariance of deformations which are related to pathology, changes in view geometry, or transforms in intensity. Atlas-based methods are more robust and used for segmentation based on probabilistic atlases [3-5]. Supervised, discriminative classification approaches have also been successful [6-8].
In atlas-based methods, spatial information is transferred from an existing dataset (labeled atlas) to a previously unseen context (target) through registration. Several studies have shown that multiple atlases (i.e., multi-atlas segmentation) outperform schemes that use a single atlas [9, 10]. Voting strategies including majority vote [9, 11], weighted vote [12], and statistical fusion (e.g., Simultaneous Truth and Performance Level Estimation, STAPLE [13]) provide structured methods for combing potentially conflicting labels from multiple atlases. A critical assumption of atlas-based methods is that it is possible to find a homology between atlases and the target. Registration of the abdominal CT volumes is known to be problematic. General purpose image-registration methods have been reported to have limited success aligning the spleen based on clinically acquired data.

Recently, Criminisi and his collaborators advocated a different concept of “segmentation” for abdominal organ localization. Rather than estimate voxelwise membership functions, they estimate bounding boxes for organs without regards to precise localization [14] through the use of random forest classification [15-17]. In [14], Criminisi et al localized ten anatomical structures including heart, liver, and spleen with regression forests and achieved better accuracy than atlas-based techniques [18].

Herein, we combine the bounding box localization method, which we refer to as regression forest region recognition (RFRR), with a traditional multi-atlas framework. We focus on spleen segmentation by first constructing RFRR classifiers for the bounding boxes for the spleen and then applying multi-atlas label fusion to perform accurate segmentation as illustrated in Fig. 1.

This manuscript is organized as follows. Section 2 details the construction of region recognition bounding box for spleen region and multi-atlas label fusion. Section 3 presents the experiments and results quantitatively. Section 4 concludes with a brief discussion.

2 Theory

We use RRFR to detect regions of interest and apply multi-atlas label fusion on region focused images to perform accurate spleen segmentation.

Briefly, T trees are created to form a classification forest for an approximate spleen bounding volume. To create each tree, a random sample of the training volumes is selected. All voxels within the selected volumes are used to build a tree. Each training voxel, \( v_i \), consists of its own position, \( v_i \), and the relative distance between it and each box, denoted as \( d_v^j \). All data are used to design the tree, starting at the root, flowing through the chain of internal nodes, and finally to a single leaf node for each input point. At each level, a “feature response” operator is created to maximize the information gain at that level for the selected training data. At each leaf, a regression model is trained to predict the distance to each bounding box with a Gaussian distribution model. Conceptually, the set of all distributions captured at the leaves for a single source image forms a posterior distribution that links the input (image data) and output (bounding box prediction) spaces. Note that the multiple, albeit simple, leaves partition the input/output space with complex non-linear mapping.
Each tree node is a binary decision that splits input data into “left” or “right” groups. To construct a tree, for each non-leaf node, a single feature operator and threshold pair are chosen to split the voxels coming into two subsets. To train a node, we randomly choose $k$ feature operators $\theta_1, \ldots, \theta_k$. For each operator, we calculate the corresponding feature response for all training voxels and split the voxels to compute the information gain based on the partition. During this training process, the chosen $\theta_j$ and the specific threshold $t_{\theta_j}$ are stored in the node and used for partitioning the unseen testing data in the future. The tree stops growing if its depth reaches a specified threshold or the information gain calculated is lower than a predefined value.

The training procedure maximizes the information gain of the splitting function by minimizing the determinants of the covariance matrix associated with each region of interest (ROI, i.e., the collection of spatial locations of the input points at a particular node). In this way, the spatial uncertainty (in the source space) in the probabilistic vote cast by each leaf cluster of voxels on each ROI decreases. Since voxels within one node share similar Gaussian distribution of $d_1$ and feature pattern, nodes with lower covariance demonstrate higher confidence for predicting the corresponding ROI. The deeper one tree grows, the smaller the leaf node clusters get, so the feature patterns are more precise and the incremental information gain is lower.

To combat the problem of over-fitting, multiple trees are trained on different random partitions of the training data. The process is repeated for each tree in the forest.

### 2.1.1 Feature Response Operators

The role of feature response function is to split input voxels within a node based on their appearance and the context of the extended image neighborhood to increase the accuracy with which one can predict the position and size of all ROIs at a leaf node.
To compute the feature response for each voxel, we choose a feature operator \( \theta_i \), which contains a reference point \( R_i \) and 3-D region \( F_i \). The region is selected to be a rectangular cuboid with each length selected uniformly at random between 7 and 14 mm at a random distance of 14 to 28 mm along each cardinal axis. In the Criminisi approach, for each voxel \( v_i \), the feature response is calculated as

\[
f(v_i, \theta_i) = \frac{1}{|F_i|} \sum_{q \in F_i} I(q)
\]

where \( F_i \) is a 3-D region from \( v_i \), referenced to the relationship between \( R_i \) and \( F_i \). This appearance indicates intensity in the neighborhood of \( v_i \), referenced to the operator information. In this way, voxels sharing similar intensity appearance cluster together and predict these regions together with their own contribution weight.

Here, we use both the original class of features and introduce a second class of features which are centered on the voxel of interest. For each voxel \( v_i \), we choose three square regions centered at the voxels with fixed lengths of \( l_1, l_2, \) and \( l_3 \), where \( l_1 = 1 \) mm, \( l_2 = 2 \) mm, and \( l_3 = 3 \) mm. The local shape feature response for voxel \( v_i \) is defined as

\[
f(v_i, \theta_i) = \sum_{k=1}^{3} \alpha_k |F_{ik}|^{-1} \sum_{q \in F_{ik}} I(q)
\]

where \( \alpha_1 = 0.6, \alpha_2 = 0.3, \) and \( \alpha_3 = 0.1 \). These values were chosen based on low resolution preliminary data. Feature vector and parameter optimization is an area of continuing interest.

### 2.1.2 Operator Optimization

At each node, 100 feature operators were generated, 50 of each type. For each feature operator, the corresponding feature response \( f(v_i, \theta_i) \) is calculated for each training voxel, where \( \theta_i \in \{ \theta_1, \theta_2, ..., \theta_k \} \). Then for feature response, we randomly select two features \( f_1 \) and \( f_2 \) for each \( \theta_i \) and sort them to calculate the split threshold,

\[
t_{\theta_i} = f_{1i} + \alpha(f_{2i} - f_{1i})
\]

where \( f_{2i} > f_{1i} \) and \( \alpha \) is a random number between 0 and 1. The value of the split threshold is used to split all the voxels into two clusters. Initially the threshold \( t_{\theta_k} \) is set to \(-\infty\), and all the voxels are in the same cluster, which form the original entropy for the whole voxels. Then when the each operator is referenced, voxels are split into left and right set of a tree based on the threshold and split function.

According to this partition, information gain \( IG_{\theta_i} \) for choosing \( \theta_i \) is obtained by calculating the entropy of left and right children node from that of parent node. Entropy here is used to measure the purity of the probability density of the real-valued predictions. Each splitting based on \( \theta_i \) can be regarded as a representation of a kind of feature-based partition. As long as we perform the partition regressively, meaningful feature of ROIs can be captured and used to predict that region which then yields a high prediction confidence.

After training, the jth split node is associated with the feature operator \( \theta_j \) and threshold \( t_{\theta_j} \). At each leaf node, we store the learned continuous parameters mean \( \bar{d}_i \)
(d₁ ∈ {d₁, d₂, ..., d_r}) and covariance matrix $\text{Cov}(S_i^j)$ (for each ROI).

### 2.1.2 Bounding Box Estimation

To estimate a bounding box, each voxel in the unlabeled data is passed through each tree in the forest from root to leaf based on the feature operator $θ_j$ and threshold $t_{θ_k}$ pre-stored in each node during training process. Note that the tree evaluation is centered on each voxel, but uses the entire image context through the $F_j$ regions stored at each tree node. Each testing voxel is selected by the stored features and then clustered into one leaf node per tree. After sending all voxels into leaf nodes, the nodes with highest confidence (least variance) for each ROI are selected from the forest to predict position and size of the region. One leaf node may have high confidence for predicting one region (i.e., bounding box edge) while low confidence for another, so we choose nodes with confidence higher than a threshold to predict a corresponding bounding box to locating a specific ROI.

The distribution stored in each leaf is represented as $p(d_i | l) = N\left(d_i; \overline{d_i}, \text{Cov}(d_i)\right)$. Since voxels within each leaf node share similar feature patterns and distributions for $d_i$, we estimate the voxel position $\hat{v}$ and use this estimate to predict the posterior for the absolute bounding box position $b_i$ by setting $\hat{v}$ as the mean position of voxels within the same leaf node. The continuous parameters for bounding box position can be obtained as $\overline{b_i}(v) = \hat{v} - \overline{d_i}(v)$. So the posterior for the absolute bounding box position given leaf can be calculated as $p(b_i | l) = N(d_i; \overline{d_i}, \text{Cov}(d_i))$. Therefore, the posterior probability for $b_i$ can be calculated as $p(b_i) = \sum_{l \in L} p(b_i | l) p(l)$, where $L$ denotes the set of leaf nodes we selected to predict the $i$th ROI in the forest and $p(l) = \frac{1}{|L|}$. The leaf nodes are independent from which tree in the forest they come from. Then the final estimation of $b_i$ can be derived from each predicted value of subset of leaves as,

$$b_i = \int_{b_i} b_i p(b_i) \, db_i$$

### 2.2 Multi-atlas Segmentation for Spleen

We follow a traditional multi-atlas approach using modern statistical label fusion approaches. Briefly, training images are cropped according to the true manual label spleen region with 10 mm padding around all sides of the manually specified labels. A bounding box is identified for each testing image using the RFRR method and is correspondingly cropped with 10 mm of padding. Specific parameters for training were (1) maximum depth of 7 levels, (2) minimum entropy again to increase tree depth of 0.05, (3) confidence threshold for including a voxel in the prediction of 0.9, (4) MRF parameters of 0.2 ($\beta$ and compatibility matrix for transition $[0 \cdot -5; -5 0]$), and (5) . Note that the training of the RFRR method did not use the testing data. All
cropped training images are registered to each cropped testing image using the NiftyReg package [19]. The corresponding labels are transferred to the target space in each case by applying the deformation field generated by image registration. NiftyReg is a two stage registration process; first, affine registration is performed based on a block matching and trimmed least square scheme, and, second, non-rigid registration is performed using free-form deformation based on normalized mutual information and bending energy.

After registration, all the propagated labels for each target image are fused using the Non-Local Spatial STAPLE (NLSS) statistical label fusion algorithm (implementation available in http://www.nitrc.org/projects/jist). The search neighborhood is initialized to an 11x11x11 window centered at the target voxel of interest and a patch neighborhood of 1x1x1 was used also centered at the voxel of interest. The standard deviation of the Gaussian intensity difference model (which controls how selective the non-local approach is in determining the correspondence between various voxels) is set to 2.0. The standard deviation of the Gaussian distance model is set to 0.2. This weights voxels based upon their distance to the current target voxel of interest. A 5x5x5 voxels sliding window is used for region-wise performance level estimation. A Markov Random Field label prior smoothing is performed on the label fusion probability estimate.

3 Methods and Results

Twenty baseline (pre-treatment) portal venous phase contrast enhanced CT scans were randomly selected from a population of patients undergoing a randomized clinical trial for FOLFOX therapy. Images were approximately 512x512x152 with a resolution of 0.7x0.7x3mm and covered the lower thorax (from the mid-cardiac region) through the abdomen (to the pelvis). Exclusion criteria were poor timing of contrast bolus administration (not portal venous phase) or aberrant patient positioning. Spleens were manually labeled by an experienced radiologist on a volumetric basis using the MIPAV software (NIH, Bethesda, MD).

The baseline (whole-volume) multi-atlas registration procedure was performed with 4-fold cross-validation (i.e., four sets of 15 trainings atlases and 5 testing images). For each testing image, pairwise affine registration was performed using NiftyReg with the training atlases, and the resulting affine matrices were used to transfer the labels to the target space. Note that in prior work, non-rigid whole abdomen registration with NiftyReg, radial basis methods [20], and ANTS SyN [21, 22] resulted in worse overlap than affine methods. The labels were fused using the Non-Local STAPLE statistical label fusion algorithm with a search neighborhood of 5x5x5 voxels, patch neighborhood of 1x1x1 voxel and the standard deviation parameters same as above [20].

The RFRR multi-atlas method was applied with 4-fold cross validation (i.e., four sets of 15 training atlases and 5 testing images). For each cohort, an RFRR classified was trained on the training data and applied to the testing data. Training time was approximately 60 hours per tree (with 5 trees trained in parallel). Bounding box estimation time was 10 minutes per volume. Representative results are illustrated in
Fig. 2. The NLSS was the primary method used to fuse labels registered based on the localized regions of interest, described above. As additional benchmarks, we fused registered labels with majority vote and local weighted vote label fusion.

Quantitative accuracy of the automatic segmentation was evaluated on a volumetric basis using Dice similarity coefficient (DSC), symmetric mean surface distance error (MSDE), and spleen volumes. The proposed RFRR multi-atlas method provides improvement over the other approaches as seen in Fig 3. Wilcoxon signed rank tests showed that the RFRR multi-atlas method resulted in significantly higher DSC than the traditional multi-atlas method ($\Delta=0.17$, $p<0.01$). The NLSS statistical fusion method resulted in higher DSC than either majority vote ($\Delta=0.13$, $p<0.01$) and local vote ($\Delta=0.12$, $p<0.01$) when considering the RFRR registered labels. Correlation between the true volumes with the volumes of RFRR multi-atlas (NLSS) was 0.934, while the correlation with the traditional multi-atlas method was 0.650.

Using the RFRR multi-atlas approach, the median values for DSC and MSDE were 0.86 and 3.1mm, respectively. A decrease in the number of outliers is observed in terms of Dice Similarity Coefficient (DSC). Selected typical and poor subject DSC results are shown in Fig. 4. The pointwise surface distance errors for all 20 subjects are shown in Fig. 5.

4 Conclusions

This manuscript presents the RFRR multi-atlas method for spleen segmentation on clinically acquired CT volumes. This approach combines regression forest region recognition and multi-atlas label fusion techniques. For 13 of 20 subjects, MSDE was less than 4mm, and for 17 of 20, MSDE was less than 10mm. The subjects with moderate degrees of errors (4-10mm) appeared to be corrupted by confusion with sections of the kidney and splenic vessels. The three subjects with substantial errors had unusual contrast agent accumulation. To address these concerns in continuing work, we are refining the space of exemplar atlases with additional subjects.
Moreover, a multi-label framework that includes splenic vessels and left kidney could help distinguish ambiguous boundary. Finally, we are working to include shape regularization to reduce spatial outliers created during statistical fusion so improve MSDE. Nevertheless, the current RFRR approach is highly promising and could greatly assist large-scale study of splenic characteristics provided a modest degree of expert review and correction.

5 Acknowledgements

This work was supported by NIH 1R03EB012461, NIH 2R01EB006136, NIH R01EB006193, and ViSE/VICTR VR3029. The project described was supported by the National Center for Research Resources, Grant UL1 RR024975-01, and is now at the National Center for Advancing Translational Sciences, Grant 2 UL1 TR000445-06. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN.

References


Fig. 3. (A) Spleen volume comparison for manual segmentation and automatic segmentation method. (B) Dice similarity coefficient for the traditional multi-atlas segmentation method and region recognition enhanced for local weighted vote, majority vote and non-local spatial stable method. (C) Symmetric mean surface distance for the traditional multi-atlas segmentation method and region recognition enhanced for local weighted vote, majority vote and non-local spatial stable method.

Fig. 4. Automatic segmentations for selected typical and poor quality results.
Fig. 5. RFRR multi-atlas segmentation errors are shown for all 20 subjects based on the manually drawn surfaces colored by the distance (in mm) between the manual segmentation and the RFRR multi-atlas segmentation. The left/right pairs show the front/back perspectives of the same subject. Within these subjects, more than half achieved surface distance within 10mm.