Weighted Voting in 3D Random Forest Segmentation

M. Yaqub[1,2], P. Mahon[3], M. K. Javaid[1], C. Cooper[1], J. A. Noble[2]
[1] NDORMS, University of Oxford, [2] IBME, Department of Engineering Science, University of Oxford, [3] MRC Epidemiology Resource Centre, University of Southampton

Abstract. The traditional random forests technique has shown good classification accuracy for 2D object segmentation in natural images. However, the technique suffers from a few problems when extending it to 3D or 4D images which are of great interest in biomedical image analysis. In this paper, we develop an automatic 3D random forests method which is applied to segment the fetal femur in 3D ultrasound. The proposed technique trains balanced trees from imbalanced data. A weighted voting mechanism is proposed to generate the probabilistic class label. A cross validation on 20 3D fetal ultrasound volumes shows promising results. Experiments show that our technique achieves segmentation and measurements close to the accuracy of expert delineations. The method runs in a few seconds on a standard PC and hence is well-suited for clinical applications.

1 Introduction

The novelty of this work is to extend the conventional Random Forests [1] (RF) technique to provide an efficient method for 3D or 4D image segmentation. In addition, we provide a robust testing by weighting the class decision of each tree. The conventional RF technique has already been used to segment 2D images [2, 3] but great interest in medical image analysis raise the issue of having such technique to accurately and efficiently segment volumetric objects. 3D features are required to represent a 3D object of interest therefore we illustrate how to extend several features to 3D efficiently. The technique has been validated and applied to segment the fetal femur in 3D ultrasound images although our technique is equally applicable to other problems.

Manual measurements can be inaccurate, tedious and time consuming. Another major problem with manual segmentation is intra and inter-observer reproducibility. The problem becomes harder when measuring volumetric structures where the errors propagate. Therefore, there is an urgent need to automate this process, enhance reproducibility and minimize the source of errors.

Recently, learning-based techniques have been proposed for segmentation. *Random forests* [1] is a learning-based technique in which training using a gold standard segmentation is done by building multiple decision trees in which every node except the leaves is a decision node that contains a feature (this is called a variable in statistics terms) and its corresponding threshold. Every leaf node contains a probabilistic class distribution (histogram of class labels for the voxels that have reached that node). Testing is performed by traversing voxels over the trees starting from the root of each tree to a leaf node. The voxels are split at a given node depending on the classification of the feature/threshold at that node. The average probabilistic decision of the class distribution from all trees is considered the final probabilistic class distribution of the test case (voxel label in this scenario). For more information see [1-3] and Fig. 1. RF can achieve comparable accuracy to boosting while being faster [4]. In addition, randomness in 1) choosing a sample training set for each tree and 2) choosing a subset of features to try at each node provides better generalization and helps avoid over-fitting. RF has also shown to have robustness to noise and ambiguity between classes in the training data which makes the technique suitable to segment ultrasound data [1].

If an equal vote from each tree is used, the decision can be biased by the strength of the classifiers on the decision node path (see green nodes in Fig. 1). For example, if the forest has 10 trees and the first one has a max depth of 5 while the other 9 trees have a max depth of 12 then the decision made by the first tree depends on up to 5 classifiers which may provide a poor accuracy compared to the trees which have up to 12 classifiers on every path. In addition, the accuracy of each classifier affects the decision. This implies that it would be a better strategy if each tree contributes a weighted vote toward the final decision.

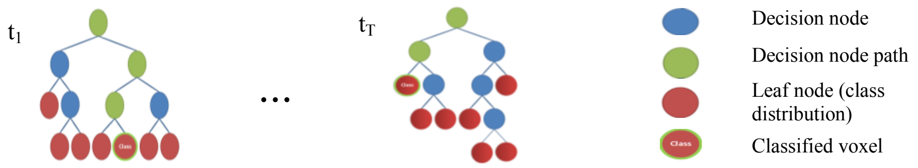


Fig. 1. Random forests which contain *T* decision trees. Decision is made as a combination of class distribution (*red circles with green outline*) from every tree (t_i).

2 Method

2.1 Problem Description

The ultimate goal of this work is to extend the traditional RF technique to 3D image segmentation and provide robust and meaningful 3D feature sets that can be computed efficiently in 3D. In addition, we provide a weighted decision that depends on the strength of the features used in each tree.

2.2 3D Feature Sets

Each node in the classification tree in the RF framework is a classifier. The classifier is in reality a feature and its threshold. In conventional RF, n' features are randomly selected out of n features in the pool (in 3D images, n can be a very huge number, e.g., 10^8). This sub-section describes how to create this feature pool.

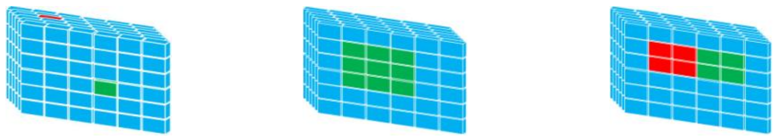
Several challenging properties of ultrasound data like shadowing, speckle and other artifacts make the problem hard. Therefore, "intelligent" features are required to capture all variations. Several feature sets are constructed for a given image. We use the phrase "feature set" to denote the group of features of the same type but with different window sizes and locations around a Voxel Of Interest (VoxOI). Unary3D, binary3D, rectangle3D [3], Haar3D [5, 6] feature sets are used and averaged rectangle3D and position3D feature sets are proposed. Fig. 2 illustrates some of these feature sets. These features are extracted from image voxels.

A unary3D feature is the intensity value of a random voxel within a random size window around VoxOI. A binary3D feature is the sum, difference or absolute difference of two random voxels in a random window around VoxOI [3]. A rectangle3D feature is the sum of all voxels of a random size rectangular cuboid starting from a random coordinate around the VoxOI. These feature sets have shown good performance in natural image segmentation [2, 3]. We extend the 2D integral images [7] to 3D, see equations (1) and (2), to find the 3D rectangular sum efficiently. Notice that the 3D integral image can be efficiently computed in one pass. Since rectangle3D features depend on the size of rectangular cuboid which is biased to its dimensions we propose an averaged rectangle3D. The averaged rectangle3D feature set is actually a unary3D feature set of the sub-sampled image in a multi-resolutions image. Haar3D features are used to capture edge regions [6]. Finally, position3D features are used to capture the spatial locations of the voxels. This feature set helps discard many regions that have similar intensity and edge information to the object of interest (e.g., in our case allowing to distinguish the femur from other structures like tibia).

$$In\ a\ 3\,D\ img\ i:\ ii(x,y,z) = \sum_{i=1}^{x}\sum_{j=1}^{y}\sum_{k=1}^{z} i(x,y,z) \quad (1)$$

$$\begin{aligned} &3\,D - rect - sum\,(x_2, y_2, z_2),(x_1, y_1, z_1) = \\ &\quad ii(x_2, y_2, z_2) + ii(x_1, y_1, z_2) + ii(x_1, y_2, z_1) + ii(x_2, y_1, z_1) \\ &\quad - ii(x_1, y_2, z_2) - ii(x_2, y_1, z_2) - ii(x_2, y_2, z_1) - ii(x_1, y_1, z_1) \end{aligned} \quad (2)$$

In each feature set many features exist. For instance, in a rectangle3D feature set 11^6 features can be generated with a maximum rectangle3D size of (11, 11, 11) starting from a random voxel within a window of a maximum size (11, 11, 11) around VoxOI. Calculating such features in 3D requires considerably more time than in 2D. In addition, many of these features are redundant and many are poor to be used for classification. Therefore, a weighted decision from each tree should give a more accurate classification.



(a) binary (b) Rectangle (c) Vertical Haar

Fig. 2. Examples of the feature sets. The *green* color voxels are summed and subtracted from the summed *red* voxels.

2.3 Training Random Forests

In the traditional RF, the training phase proceeds by building randomized decision trees. The number of trees is set before hand. A top-down construction for every tree is performed starting from the root node. Each tree is trained on a random set of the training points with replacement. For each node in the tree n' features from the feature pool are randomly selected without replacement. The "best" feature out of n' with the "best" threshold is selected as a classifier in the tree node. Information gain is usually used to decide the performance of a classifier. The training set is then divided into two sub-sets according to the results of the classifier to left and right branches. The same process is continued recursively for each sub-set until the maximum tree height is reached or no more gain is achieved. For more information see [1]. After trees construction, every leaf node contains a probabilistic class distribution $P(c_i|l)$ for each class which is the histogram of the training examples of class label c_i that reached leaf node l.

2.4 Segmentation & Measurements

In traditional RF, classifying new voxels proceeds by testing each voxel on the features/thresholds for every tree starting from the root to a leaf node. The probability for a voxel v_i to belong to a specific class c_j is the percentage of voxels of class c_j that reached the leaf node with respect to all voxels reached it (v_{leaf}) during training (3). The probabilities from all trees are averaged to generate the final probabilistic decision of a voxel v_i belonging to a class c_j (4).

$$p(c_j \mid v_i, tree_t) = \frac{hist(v_{leaf} \in c_j, l_t)}{hist(v_{leaf}, l_t)} \quad (3) \qquad p(c \mid v_i, RF) = \frac{1}{T}\sum_{t=1}^{T} p(c \mid v_i, tree_t) \quad (4)$$

Here T is the number of the trees and l_t is a leaf node at $tree_t$.

One major issue is the equal vote ($1/T$) from each tree where some trees may provide a bad classification accuracy. One solution could be to increase the number of trees but this significantly increases the training and testing time in the RF. Therefore, we propose a weighted voting in which the vote is weighted depending on the features used in each tree starting from the root until the leaf node. The decision from each tree is based on the classification accuracy of the nodes visited for every voxel v_i. To embed this into the RF framework, a weighted sum of trees probabilities is proposed and equation (4) is generalized to (5).

$$p(c \mid v_i, RF) = \sum_{t=1}^{T} \alpha_t \, p(c \mid v_i, tree_t) \quad (5) \qquad \text{where} \quad \alpha_t = \frac{\frac{1}{F}\sum_{f=1}^{F} Score_f(tree_t)}{\sum_{i=1}^{T}\left(\frac{1}{F}\sum_{f=1}^{F} Score_f(tree_i)\right)} \quad (6)$$

Here F is the total number of features on the path from the root to the leaf when classifying a voxel and $Score_f(tree_t)$ is the training score of a feature f on a path at tree t. Finally, the volume of the segmentation of class c_j is easily found by multiplying the number of segmented voxels by the voxel spacing.

2.5 Post-processing

This step is application-specific and is mainly applied here to reject regions with similar local shape and intensity distribution to the object of interest. Although RF provides good classification accuracy, it is a discriminative model that captures local similarities. As a result, any structure which looks similar in intensity distribution and local shape can be regarded as the object of interest. A position feature set is added to reject such regions. Unfortunately, in our specific application some femur like structures are close to the femur and therefore position features may not be able to distinguish between the two (e.g., the femur is connected to tibia via the knee ligaments). To accommodate this, the largest 3D connected component was automatically selected (the femur).

3 Experimental Results

Several measurements of fetal structures from 2D ultrasound images are important to diagnose the growth of the fetus and estimate gestational age and birth weight [8, 9]. Clinicians usually measure head circumference, biparietal diameter, abdominal circumference and femur length. Several research groups have studied and manually measured the fetal femur [8-10]. They have mainly focused on measuring femur length to correlate it with gestational age or birth weight. Several research groups have tried to automate the process of segmenting and measuring such structures in 2D ultrasound images [5, 11]. To our knowledge we are the first to investigate the problem of automatic femur volume segmentation in 3D ultrasound images.

3.1 Dataset

We tested our technique on 20 3D ultrasound volumes [9] acquired on 19 weeks fetuses ±6 days using a GE Voluson 730 scanner. Volumes dimensions are approximately 70×70×140 with a (0.5×0.5×0.5) mm^3 voxel spacing. Although out-of-bag error estimate can be used as a classification error measure [1], cross validation was performed on the 20 volumes by using 18 images for training and two for testing. Cross validation provides a more general and realistic error measure compared to the out-of-bag error in this application.

3.2 Validation methodology

Experiments on the traditional and weighted RF are reported to support the proposed technique. RF requires several parameters to be set. The parameters were fixed for all experiments (T = 10, max-tree-depth = 10, n' = 100). Recall and precision were calculated to measure how well the segmentation of the proposed technique compared to an expert manual segmentation according to (7) and (8) respectively.

$$\text{Recall} = \frac{TP}{TP + FP} \quad (7) \qquad \text{Precision} = \frac{TP}{TP + FN} \quad (8)$$

Where: TP is True Positive
FP is False Positive
FN is False Negative

Recall and precision comparisons of the 20 volumes for the traditional RF and the weighted RF are shown in Table 1. Notice that the higher the recall the closer the segmentation is to the ground truth. Bland-Altman plots for the volume measurements to compare the manual segmentation and the traditional and weighted RF techniques show that the weighted RF has the minimum bias and tightest standard deviation bounds (Fig. 3). Visual comparisons between the manual segmentation and the both RF methods are shown in Fig. 4.

The training and segmentation times for the RF technique are shown in Table 2. These times are for one experiment where 18 ultrasound images were used for training and one for testing, T=10, n'=100, Max-tree-depth=10, n~=8*10^6.

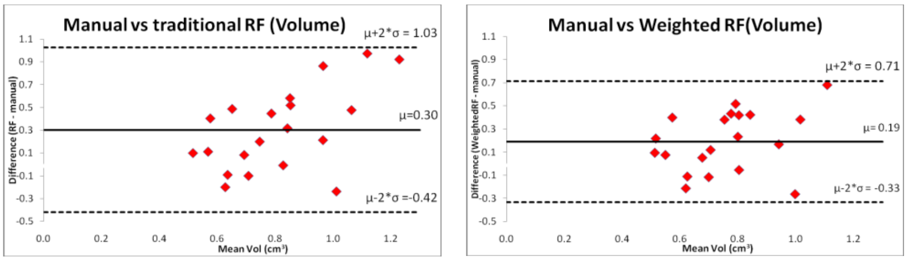


Fig. 3. Bland-Altman plots for the segmented femur volumes. Left: manual vs. traditional. RF. Right: manual vs. weighted RF.

Table 1. Recall & precision for the traditional and weighted RF methods.

	μ±σ Recall	μ±σ Precision
Trad. RF	64%±18%	88%±11%
Weighted RF	70%±15%	88%±11%

Table 2. Training & test time for traditional and weighted RF methods.

	Training time (Hours)	Segmentation time (Sec)
Trad. RF	18	13
Weighted RF	18	22

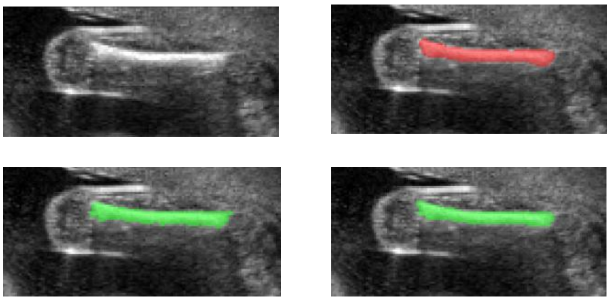


Fig. 4. A 2D slice of the segmentation using 3D random forests. Top is an original longitudinal (left) and ground truth (right). Bottom is the segmentation using traditional RF (left) and weighted RF (right).

4 Conclusions & Future work

In this paper, the RF technique has been extended from the traditional 2D RF to 3D. We have shown that using weighted class decision from each tree in RF outperforms the conventional method. The technique has shown good accuracy and performance on the problem of fetal femur segmentation in 3D ultrasound data. Validation has been performed on a good size dataset which showed promising results. One major issue to consider is to eliminate irrelevant features in the huge feature pool. This will theoretically provide better classification accuracy. A second issue is how to integrate global shape information in the RF framework since RF mainly capture the local shape information of the object of interest. Researchers have looked into this issue by applying a generative model to the results of the discriminative model (e.g., Boosting and its variations, RF, etc...) [12]. Specific to our application, the feature set could also be extended to account for the signal attenuation for both the distal and proximal ends of the femur. We also plan to study the intra and inter-observer reproducibility by doing multiple manual segmentations from multiple experts. Finally, this approach is general and not restricted to the femur or indeed ultrasound. We plan to look at other applications too.

References

[1] L. Breiman, “Random Forests,” *Machine Learning,* vol. 45(1), pp. 5-32, 2001.

[2] F. Schroff *et al.*, “Object Class Segmentation using Random Forests,” in BMVC, 2008.

[3] J. Shotton *et al.*, “Semantic Texton Forests for Image Categorization and Segmentation,” in CVPR, 2008.

[4] R. Caruana, and A. Niculescu-Mizil, “An Empirical Comparison of Supervised Learning Algorithms,” in ICML, 2006.

[5] G. Carneiro *et al.*, “Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree,” *IEEE TMI,* vol. 27(9), pp. 1342-55, 2008.

[6] M. Oren *et al.*, “Pedestrian detection using wavelet templates,” in IEEE CVPR, 1997, pp. 193-199.

[7] P. Viola, and M. Jones, “Robust Real Time Object Detection,” *IJCV*, 2001.

[8] L. S. Chitty, and D. G. Altman, “Charts of fetal size: limb bones,” *British Journal of Obstetrics and Gynaecology,* vol. 109, pp. 919–929, 2002.

[9] P. A. Mahon, “Ultrasound assessment of fetal musculo-skeletal development,” PhD. Thesis, University of Southampton, Southampton, 2007,PhD Thesis.

[10] C.-H. Chang *et al.*, “Prenatal Detection of Fetal Growth Restriction by Fetal Femur Volume: Efficacy Assessment Using Three-Dimensional Ultrasound ” *UMB,* vol. 33(3), pp. 335-341, 2007.

[11] S. M. Jardim, and M. A. Figueiredo, “Segmentation of fetal ultrasound images,” *UMB,* vol. 31(2), pp. 243-50, 2005.

[12] Z. Tu *et al.*, “Brain anatomical structure segmentation by hybrid discriminative/generative models,” *IEEE TMI,* vol. 27(4), pp. 495-508, 2008.