# Learning from Multiple Experts with Random Forests: Application to the Segmentation of the Midbrain in 3D Ultrasound

Pierre Chatelain[1,2], Olivier Pauly[1,3], Loïc Peter[1], Seyed-Ahmad Ahmadi[1], Annika Plate[4], Kai Bötzel[4], and Nassir Navab[1]

[1] Computer Aided Medical Procedures, Technische Universität München, Germany
pierre.chatelain@cs.tum.edu
[2] Ecole Normale Supérieure de Cachan, Antenne de Bretagne, France
[3] Institute of Biomathematics and Biometry, Helmholtz Zentrum München, Germany
[4] Department of Neurology, Ludwig-Maximilians-Universität München, Germany

**Abstract.** In the field of computer aided medical image analysis, it is often difficult to obtain reliable ground truth for evaluating algorithms or supervising statistical learning procedures. In this paper we present a new method for training a classification forest from images labelled by variably performing experts, while simultaneously evaluating the performance of each expert. Our approach builds upon state-of-the-art randomized classification forest techniques for medical image segmentation and recent methods for the fusion of multiple expert decisions. By incorporating the performance evaluation within the training phase, we obtain a novel forest framework for learning from conflicting expert decisions, accounting for both inter- and intra-expert variability. We demonstrate on a synthetic example that our method allows to retrieve the correct segmentation among other incorrectly labelled images, and we present an application to the automatic segmentation of the midbrain in 3D transcranial ultrasound images.

## 1 Introduction

Manual segmentation of medical images is often time-consuming and highly subjective, suffering from both inter-observer and intra-observer performance variability, due to differences of interpretation and level of expertise. The segmentation task is especially challenging in ultrasound images, whose low signal-to-noise ratio leaves a large place to interpretation. As a precise segmentation is crucial for the analysis of images in computer-assisted diagnosis (CAD), it is important to take into account this variability of performance when fusing the segmentations given by different experts. Moreover, the lack of reliable ground truth in medical data limits the use of supervised learning methods to perform automatic segmentation. These methods therefore have to be adapted to exploit efficiently a multiplicity of labellings of variable accuracy.

The simplest way to make an estimate of the ground truth from multiple segmentations is to perform a majority voting among the experts. However,

K. Mori et al. (Eds.): MICCAI 2013, Part II, LNCS 8150, pp. 230–237, 2013.

this method assumes that all experts are equally good, and fails as soon as there are more novice segmenters than good experts. Warfield et al. [1] propose a method to perform simultaneous truth and performance level estimation (STAPLE), consisting in finding the performance parameters (sensitivity and specificity) which maximize the data likelihood. This optimization is performed using the expectation-maximization (EM) algorithm. They also propose to incorporate spatial constraints with a Markov random field (MRF) model [1,2] to obtain a spatially consistent estimation of the ground truth. This method has been widely accepted and successfully used for ground truth estimation or validation of segmentation algorithms [3]. However, the global model for experts' performances does not take into account the intra-observer variations of performance level that are frequent in medical image segmentation. For instance, an expert can be good at segmenting certain parts of the object of interest, while classifying wrongly other parts of the same object. Commowick et al. [4] recently addressed this issue by proposing an adaptation of STAPLE, which estimates spatially varying performance levels, using a sliding window technique.

Although these algorithms are very efficient for label fusion, they lack a link to the images which have been segmented. Indeed, the fusion of labels is based only on the segmentations independently from the original images, and thus can not assess the visual coherence of each expert's labelling. However, understanding the link between the segmentations and the image is crucial if one wants to train a statistical classifier in order to automatically segment new images. Raykar et al. [5] introduce a general framework for supervised learning from multiple annotators, which overcomes this issue by performing jointly the performance evaluation and the learning. However, this approach assumes that the performance levels are not dependent on the instance on which the experts make the decision. This assumption is in reality not true for the task of medical image segmentation, where there is also a strong intra-expert performance variability. The accuracy of an expert can for instance vary significantly from an image to another, depending on the image quality, the expert's experience and tiredness, and the time allowed to perform the diagnosis.

Recently, randomized classification forests [6] have been shown to be able to capture discriminative visual features from local and/or long-range context for the analysis of medical images [7,8]. In this paper we propose a new method to train a classification forest from multi-supervised data, while jointly evaluating the performances of the experts. Our approach aims at finding which expert decisions are the most coherent with respect to the image. To this end we define a consistency measure based on the information gain, to evaluate how well the decisions of each expert can separate the chosen feature space. The motivation for this approach is that our main objective is to select relevant features to train the classifier, and not to estimate a consensus ground truth. We demonstrate the theoretical capabilities of our algorithm on a synthetic example, and validate its performance for the segmentation of medical images on real 3D transcranial ultrasound images.

## 2 Methods

Let us consider a multi-supervised learning scenario, where we are given a training set $S = \{(x_n, y_n^1, \ldots, y_n^R)\}_{n=1}^N$ of $N$ instances $x_n$ in a feature space $\mathcal{X}$ and the corresponding labels $y_n^r \in \mathcal{Y}$ provided by $R$ experts. Considering the instances and labels as realizations of two random variables $X$ and $Y$ respectively, the goal of supervised learning is to provide an approximation of the posterior distribution $\Pr(Y|X)$ in order to label previously unseen data. When the labels are provided concurrently by several experts ($R > 1$), this task requires to make an estimate of the true label $y_n$ for each instance. To this end, we define a matrix $\mathbf{Q}_{N\times R}(n, r) = (q_n^r)_{n,r}$ representing the quality of each expert decision, from which we make a probabilistic estimate $w_n = p(y_n = 1|S, \mathbf{Q})$ of the true labels.

### 2.1 Random Decision Forests

We start by shortly presenting the classic random forest (RF) framework when non-ambiguous labels are provided by a single expert ($R = 1$). Constructed as an ensemble of decorrelated decision trees, random forests [6] are a state-of-the-art learning algorithm that has been used in a wide range of applications [9]. The general idea of random forest training is to select relevant subsets of features out of $\mathcal{X}$ in order to build a piece-wise approximation of $\Pr(Y|X)$. In this paper we focus on binary classification forests ($\mathcal{Y} = \{0, 1\}$), but similar techniques can be easily applied to multi-class decision forests or regression forests.

A binary decision tree is a hierarchical collection of nodes and leaves. While each node contains a weak classifier which separates the data into two subsets of lower entropy, each leaf provides a local estimate of the class distribution. A tree is trained by optimizing recursively the parameters of each node. The entire training set is input to the root, and split across the tree according to the rules explained below. Training a node $j$ on $S_j \subset S$ consists in finding the parameters of the weak classifier that maximize the information gain (IG) of splitting $S_j$ into $S_k$ and $S_l$, defined as:

$$\mathrm{IG}(S_j, S_k, S_l) = H(S_j) - \frac{|S_k|}{|S_j|}H(S_k) - \frac{|S_l|}{|S_j|}H(S_l) \tag{1}$$

where $H(S_i)$, $i \in \{j, k, l\}$ is the empiric entropy of $S_i$. Once the weak classifier has been optimized, its parameters are stored in the node, and the training data is split accordingly into $S_k$ and $S_l$, respectively sent to the left and right child. The splitting stops when we reach a predefined maximal depth, or when the training subset does not contain enough samples. In this case, a leaf is created that stores the empiric class posterior distribution estimated from this subset.

Using a collection of decorrelated decision trees allows to increase the generalization power compared to individual trees, often suffering from overfitting. The randomness is introduced both by training each tree on a random subset of the whole training set (bagging), and by optimizing each node over a random subspace of the feature parameter space. At testing time, the output of the forest is defined as the average of the probabilistic predictions of the $T$ trees.

### 2.2 Multi-supervised Learning

Let us now consider the problem of multi-supervised learning ($R > 1$). Our main objective is to select features that are coherent with respect to both the expert decisions and the visual context, in order to build an approximation of the posterior $\Pr(Y|X)$ that generalizes well to new data instances. Therefore, in contrast to the usual label fusion framework [1,5], we evaluate the consistency of the experts with respect to the visual features, rather than their accuracy with respect to a ground truth estimate. Using the hierarchical structure of the trees, we also capture the intra-expert performance variability by selecting at each node the most consistent expert for splitting.

To this end, we define for each expert $r \in [|1, R|]$ an estimator $\widehat{E}_j^r$ of the expectation of the information gain on the training set $S_j$ sent to the node $j$:

$$\widehat{E}_j^r = \frac{1}{|\Theta_j|} \sum_{\theta \in \Theta_j} \mathrm{IG}_r(S_j, S_k(\theta), S_l(\theta)) \tag{2}$$

where the information gain $\mathrm{IG}_r(S_j, S_k(\theta), S_l(\theta))$ is estimated by (1) according to the labels of the expert $r$, and $\Theta_j$ is a randomly selected subset of the feature parameter space. Intuitively, this estimator measures how well the data can be separated according to the labels of each expert. However, it suffers from two weaknesses in lower nodes of the tree: (i) it is evaluated from less samples, and hence becomes less reliable, and (ii) it quantifies only the experts' local consistency, without considering global consistency measures. Therefore we propose to define the performance level $q_j^r$ of each expert as a linear combination of the estimators $\widehat{E}_i^r$ on the path $root = i_0(j) \rightarrow \ldots \rightarrow i_D(j) = j$ from the root to $j$:

$$q_j^r = \frac{\sum_{d=0}^{D(j)} |S_d| \widehat{E}_{i_d(j)}^r}{\sum_{d=0}^{D(j)} |S_d|} \tag{3}$$

By weighting the estimators in proportion to the size of the training subset, we give more importance to the global estimates of the experts' consistencies, but still take into account their feature-specific performances. Once the parameters $q_j^r$ have been computed, we select the expert $\hat{r}$ that is the most consistent, i.e. with the highest performance level:

$$\hat{r} = \arg\max_r q_j^r \tag{4}$$

In the considered node, we then split the data according to the feature parameters that maximize the information gain for the selected expert $\hat{r}$, thus following the decision of the most consistent expert:

$$\hat{\theta}_j = \arg\max_{\theta \in \Theta_j} = \mathrm{IG}_{\hat{r}}\left(S_j, S_k(\theta), S_l(\theta)\right) \tag{5}$$

If $j$ is a leaf with parent node $i$, we define the estimate of the ground truth as the weighted average of the expert decisions:

$$\forall n : (x_n, y_n^1, \ldots y_n^R) \in S_j,\ w_n = \frac{1}{\sum_{r=1}^R q_i^r} \sum_{r=1}^R q_i^r y_n^r \tag{6}$$

Equation (6) provides a soft voxel-wise estimate of the ground truth for each tree, and a feature-specific estimate of the posterior $\Pr(Y|X)$ in each leaf. The posterior is stored for prediction, and the ground truth estimate is averaged over the forest, thus increasing its robustness.

## 3 Experiments

We first use a synthetic example to validate our algorithm, and then demonstrate its performance on a real dataset, for the segmentation of the midbrain in 3D transcranial ultrasound (TC-US) images. For the validation we compare our multi-supervised forest to a classic random forest trained on the precomputed truth estimate of STAPLE (STAPLE+RF). We show that our estimation of the ground truth is more robust to noisy labels, and that the prediction on new images achieves an accuracy comparable to that of STAPLE+RF.

For both experiments we use the so-called context features [10], that are parameterized by an offset $\mathbf{t} \in \mathbb{R}^3$ and two cube dimensions $S^{(1)}$, $S^{(2)} \in \mathbb{R}^3$. The corresponding feature for a voxel $\mathbf{v} \in \mathbb{R}^3$ is computed as the binary difference between the mean intensity in the cube of size $S^{(1)}$ centered on $\mathbf{v}$ and the cube of size $S^{(2)}$ centered on $\mathbf{v} + \mathbf{t}$. This type of feature, already applied for MR images [10], is also well suited for ultrasound images, as it is independent of the contrast and illumination.

### 3.1 Validation on a Synthetic Example

In this synthetic example the foreground consists in cubes of various dimensions, placed randomly within the images. Gray-scale images are generated from this ground truth by defining for each image a maximal foreground intensity $\mathcal{I}_1^{max}$ and a maximal background intensity $\mathcal{I}_0^{max} < \mathcal{I}_1^{max}$. The intensity of each foreground (resp. background) voxel is drawn uniformly between 0 and $\mathcal{I}_1^{max}$ (resp. $\mathcal{I}_0^{max}$). With these settings, relying on voxelwise intensities would yield poor results, but the foreground is separable using the box features described above.

Each expert $r = 1, \ldots, R$ is modeled by a maximal offset $\mathbf{T}_r \in \mathbb{R}^3$ and three deformation parameters $(\Delta_x, \Delta_y, \Delta_z) \in \mathbb{R}^3$. For each image, given the true foreground cube centered on $\mathbf{v}_0$ and of size $(s_x, s_y, s_z)$, the segmentation of expert $r$ is a cube centered on $\mathbf{v}_0 + \frac{t_r}{\|\mathbf{T}_r\|}\mathbf{T}_r$ and of size $(s_x + \delta_x, s_y + \delta_y, s_z + \delta_z)$, where $t_r$ is drawn uniformly between 0 and $\|\mathbf{T}_r\|$, and $\delta_x$, $\delta_y$, $\delta_z$ are drawn uniformly between 0 and $\Delta_x$, $\Delta_y$, $\Delta_z$ respectively. With this model we can thus generate a great variability of expert performances.

Table 1 summarizes the results obtained by performing leave-one-out cross-validation on a set of 10 images labelled by 5 experts. We present the maximum, minimum, mean, median, and standard deviation of the F-measure, defined as $F = \frac{2 \times precision \times recall}{precision + recall}$. The multi-supervised forest and the standard forest were trained with the same parameters (16 trees, 50 tests per node, maximal depth 5, and bagging proportion of 0.5), and initialized with the same random seeds, in order to make the comparison meaningful. Fig. 1 shows an example of results

with these two methods. The estimation of the ground truth using the multi-supervised forest is in this situation clearly more robust than STAPLE. This demonstrates that using image information to assess the performance of the experts can improve the label fusion. Moreover, the estimated segmentation is in most cases closer to the actual ground truth than the expert segmentation. The prediction of the multi-supervised forest is also more accurate than the prediction of STAPLE+RF.

**Table 1.** Synthetic example – F-measure of the truth estimate and the prediction for our multi-supervised forest algorithm and STAPLE+RF

| | | Truth estimation | | Prediction | |
|---|---|---|---|---|---|
| | Experts | Our algorithm | STAPLE | Our algorithm | STAPLE+RF |
| max | 0.83 | 0.85 | 0.72 | 0.87 | 0.77 |
| mean | 0.72 | 0.74 | 0.34 | 0.74 | 0.66 |
| min | 0.67 | 0.62 | 0.11 | 0.43 | 0.35 |
| median | 0.70 | 0.74 | 0.27 | 0.82 | 0.69 |
| std | 0.06 | 0.08 | 0.19 | 0.15 | 0.13 |

(a) (b) (c) (d) (e)

**Fig. 1.** Synthetic example (a) Ground truth (b) Truth estimate of the multi-supervised forest (c) STAPLE truth estimate (d) Segmentation of the multi-supervised forest (e) Segmentation of STAPLE+RF

### 3.2 Segmentation of the Midbrain

We now demonstrate the performance of our algorithm on medical data. We use a dataset made available by the authors of [8], consisting of 22 three-dimensional TC-US images, where the midbrain has been segmented by three different experts (Fig. 2). Accurate segmentation of the midbrain is of interest for the diagnosis of Parkinson's disease, which can be performed by volumetric analysis of the substantia nigra within the midbrain [11]. This dataset is particularly challenging due to the very low signal-to-noise ratio of TC-US images. To assess the performance of our approach compared to STAPLE+RF, we perform a leave-one-out cross-validation. The forests were trained with 32 trees, a maximal depth of 20, a bagging proportion of 0.5, and 100 tests per node. In the results presented in Table 2, the multi-supervised forest provided the same overall ranking between the experts than STAPLE. However the forest ground truth

estimate was in average closer to the manual segmentations than the STAPLE estimate, which often favored one of the segmenters. The predictive results of the multi-supervised forest almost match those of STAPLE+RF. Note that we only present here the rough posteriors of the algorithms without post-processing, as the final segmentation is beyond the scope of this paper.

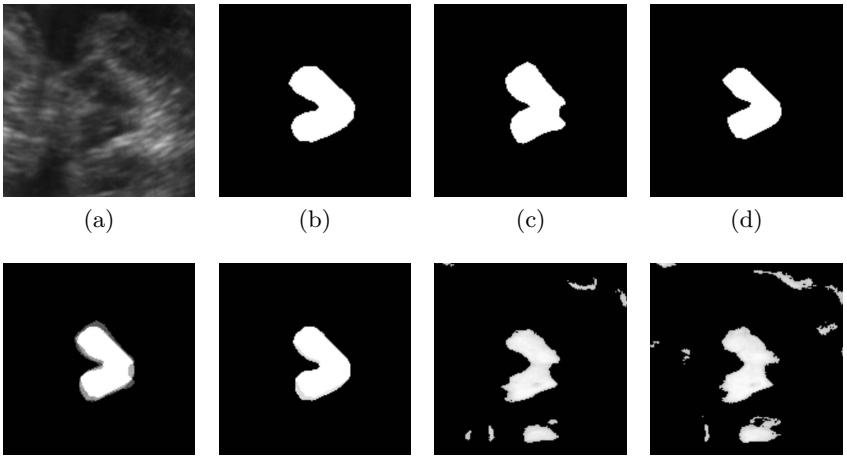


**Fig. 2.** Segmentation of the midbrain in 3D TC-US (a) Ultrasound image (b-d) Manual segmentations (e) Truth estimate of the multi-supervised forest (f) STAPLE truth estimate (g) Segmentation of the multi-supervised forest (h) Segmentation of the forest trained on STAPLE

**Table 2.** Midbrain dataset – F-measure of the truth estimate and the prediction for our multi-supervised forest algorithm and Staple+RF

| | Truth estimation | | | | | | Prediction | |
|---|---|---|---|---|---|---|---|---|
| | Our algorithm | | | STAPLE | | | Our | STAPLE |
| | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 1 | Exp. 2 | Exp. 3 | algorithm | +RF |
| max | 0.99 | 0.98 | 0.98 | 1.00 | 1.00 | 1.00 | 0.58 | 0.62 |
| mean | 0.92 | 0.96 | 0.92 | 0.89 | 0.96 | 0.90 | 0.31 | 0.33 |
| min | 0.82 | 0.89 | 0.80 | 0.81 | 0.75 | 0.71 | 0.06 | 0.05 |
| median | 0.93 | 0.97 | 0.94 | 0.89 | 1.00 | 0.91 | 0.32 | 0.35 |
| std | 0.04 | 0.02 | 0.05 | 0.04 | 0.06 | 0.07 | 0.15 | 0.17 |

## 4 Conclusion

We have proposed a new method to supervise the training of a random forest from multiple experts, that accounts for both inter- and intra-expert performance

variability. The robustness of our multi-supervised forest training algorithm was demonstrated on a synthetic example, and we showed consistent results for the segmentation of the midbrain in 3D transcranial ultrasound images. While our algorithm was developed here for binary classification, it can be naturally applied to multi-class learning or regression. As future work we wish to extend this method to label propagation, in a situation where only a few images in the training set are correctly labelled.

## References


1. 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 Transactions on Medical Imaging 23, 903–921 (2004)
2. Commowick, O., Warfield, S.K.: Incorporating priors on expert performance parameters for segmentation validation and label fusion: A maximum a posteriori STAPLE. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part III. LNCS, vol. 6363, pp. 25–32. Springer, Heidelberg (2010)
3. Martin-Fernandez, M., Bouix, S., Ungar, L., McCarley, R.W., Shenton, M.E.: Two methods for validating brain tissue classifiers. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3749, pp. 515–522. Springer, Heidelberg (2005)
4. Commowick, O., Akhondi-Asl, A., Warfield, S.: Estimating a reference standard segmentation with spatially varying performance parameters: Local MAP STAPLE. IEEE Transactions on Medical Imaging 31(8), 1593–1606 (2012)
5. Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. Journal of Machine Learning Research 11, 1297–1322 (2010)
6. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
7. Criminisi, A., Shotton, J., Bucciarelli, S.: Decision forests with long-range spatial context for organ localization in CT volumes. In: MICCAI Workshop on Probabilistic Models for Medical Image Analysis (2009)
8. Pauly, O., Ahmadi, S.-A., Plate, A., Boetzel, K., Navab, N.: Detection of substantia nigra echogenicities in 3D transcranial ultrasound for early diagnosis of parkinson disease. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 443–450. Springer, Heidelberg (2012)
9. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Foundations and Trends in Computer Graphics and Vision 7(2-3), 81–227 (2012)
10. Pauly, O., Glocker, B., Criminisi, A., Mateus, D., Martinez-Moeller, A., Nekolla, S., Navab, N.: Fast multiple organs detection and localization in whole-body MR dixon sequences. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part III. LNCS, vol. 6893, pp. 239–247. Springer, Heidelberg (2011)
11. Geng, D.Y., Li, Y.X., Zee, C.S.: Magnetic resonance imaging-based volumetric analysis of basal ganglia nuclei and substantia nigra in patients with parkinson's disease. Neurosurgery 58(2), 256–262 (2006)