Joint Classification-Regression Forests for Spatially Structured Multi-Object Segmentation

Ben Glocker[1], Olivier Pauly[2,3], Ender Konukoglu[1], Antonio Criminisi[1]

[1] Microsoft Research, Cambridge, UK
[2] Institute of Biomathematics and Biometry, Helmholtz Zentrum München, Germany
[3] Computer Aided Medical Procedures, Technische Universität München, Germany

Abstract. In many segmentation scenarios, labeled images contain rich structural information about spatial arrangement and shapes of the objects. Integrating this rich information into supervised learning techniques is promising as it generates models which go beyond learning class association, only. This paper proposes a new supervised forest model for joint classification-regression which exploits both class and structural information. Training our model is achieved by optimizing a joint objective function of pixel classification and shape regression. Shapes are represented implicitly via signed distance maps obtained directly from ground truth label maps. Thus, we can associate each image point not only with its class label, but also with its distances to object boundaries, and this at no additional cost regarding annotations. The regression component acts as spatial regularization learned from data and yields a predictor with both class and spatial consistency. In the challenging context of simultaneous multi-organ segmentation, we demonstrate the potential of our approach through experimental validation on a large dataset of 80 three-dimensional CT scans.

1 Introduction

Semantic image segmentation consists of assigning a categorical label to each pixel in an image. A common approach is to cast segmentation as a multi-label classification problem and employ a classification algorithm. In this context, supervised learning techniques have gained increased interest. Relying on the availability of annotated data, they permit to learn the relationship between visual features of pixels and their class labels during their training phase. Given an unseen image, the learned classifier is then able to predict the correct label assignment for each pixel.

Decision forests have emerged as a promising, flexible model for image understanding [1–4]. In particular, classification and regression forests have shown great performance in the tasks of supervised classification and regression such as human pose estimation [5], recognition [6], localization [7], or classification [8, 9]. Classification forests are popular because they are probabilistic and efficient, and naturally handle multi-class problems. Moreover, they often compare favorably with respect to other techniques [10, 11].

In their original implementation classification forests provide as output a class posterior distribution for each pixel independently. Recent work has started to investigate new and more complex models of structured-output forests to enable spatially consistent predictions [12–15]. However, accessible structural information about the shapes and spatial arrangement of objects present in ground truth annotations, *i.e.* label maps, is not fully exploited in previous approaches.

The main contribution of this paper is a novel joint classification-regression formulation based on decision forests which incorporates this extra information. In each tree, we learn a discrete-continuous predictor based on class *and* spatial consistency by extracting structural information from label maps. The key innovation within our approach is a simple yet elegant modification of the training objective function which enables joint learning of classification and regression. We employ signed distance maps (SDMs) in a regression objective as efficient representations of information about shapes and spatial arrangement.

Similar to pictorial structures [16] our model is particularly suited for images with multiple objects whose organization shows some consistency (*e.g.* facial features, limbs in a human body, internal organs in medical scans, etc.).

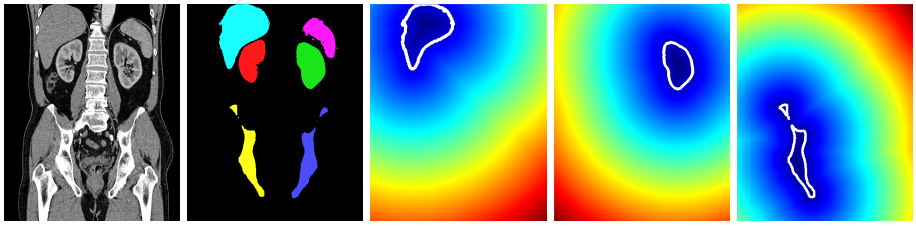
Classification and regression have been combined before in the context of decision forests for body joint prediction [17] and object detection [18]. Both approaches are quite different to ours. In [17], the prediction model is a single continuous regressor, for which training is performed *either* based on a classification or regression objective function. In [18], the training objective alternates between classification and regression, but is not based on a joint objective function.

There are many other methods which aim at solving the problem of structured multi-object segmentation. Active shape and appearance models [19], or random fields [20] are among the most successful ones. A comparison with these methods is beyond the scope of this paper. Here, we focus on one particular approach based on classification forests, and demonstrate how performance can be substantially improved through simple modifications. We believe that an isolated view on this particular modification yields more insights than a broader comparison with substantially different methods. Further, we believe that our proposed modifications can be easily integrated in existing, more complex approaches.

Experimental validation of our model is carried out on multi-organ segmentation on a challenging labeled dataset of 3D medical CT scans of 80 patients.

2 Classification-Regression Forests

In the following, we will derive a general formulation for joint classification-regression in the context of decision forests. At the same time, we will provide the necessary details for our application of multi-object segmentation. We refer the interested reader to [2] for more details on forests.



(a) CT scan (b) Label map (c) SDM liver (d) SDM kidney (e) SDM pelvis

Fig. 1. An example slice of a 3D input image in (a) with ground truth label map in (b). Besides class membership, the label map contains additional information such as a distance for each pixel to all objects of interest obtained from signed distance maps as shown in (c)-(e). The zero-level is overlaid on the distance maps for clarity. Pixels inside an object have negative distances.

2.1 Decision Forests for Supervised Learning

In its most general form, the goal of supervised, discriminative learning is to obtain the posterior distribution $p(\mathbf{y}|\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^m$ is some observation represented by a *feature vector*, and $\mathbf{y} \in \mathbb{R}^n$ is the output or *prediction* variable. Learning this distribution allows us to make predictions for new (unseen) data, *e.g.* by inferring the maximum-a-posteriori (MAP) estimate $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})$.

We assume that a set of K training examples $\mathcal{S} = \{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^{K}$ is available, from which we can learn the distribution $p(\mathbf{y}|\mathbf{x})$. In image segmentation, the entity $\mathbf{x}_k$ corresponds to a collection of image features – *e.g.* intensity or textural information – extracted for an individual pixel k. The output variable is the (one-dimensional) discrete class label $\mathbf{y}_k \in \mathcal{C}$, where $\mathcal{C}$ is a finite set of labels (or objects). The aim is then to learn a predictor that determines the probability for assigning a particular class label to a pixel of a previously unseen test image.

We employ the decision forest framework which tackles the learning problem in a divide-and-conquer fashion. A decision forest is an ensemble of (probabilistic) decision trees, where each tree t yields its own distribution $p_t(\mathbf{y}|\mathbf{x})$. By iteratively subdividing the training set within the associated features space $\mathbb{R}^m$, posterior distributions can be learned "locally" on smaller training subsets. Injecting randomness into the training phase decreases the correlation between individual trees, and increases generalization (see [1] for details).

Tree testing: A (binary) decision tree is a set of two types of nodes, the *split nodes* and the *leaf nodes*. While split nodes store decision functions, leaf nodes store empirical distributions. In order to make a prediction for previously unseen data $\mathbf{x}$, we push $\mathbf{x}$ through the tree, starting at the root node. At each split node, a (binary) decision function is applied to $\mathbf{x}$, which determines whether it is sent to the left or right child node. Once the data point reaches a leaf node, we can simply read out the stored distribution $p_t(\mathbf{y}|\mathbf{x})$. The overall prediction of the

forest with T trees can be obtained by averaging the individual tree predictions:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} p_t(\mathbf{y}|\mathbf{x}) \quad . \tag{1}$$

Tree training: The role of training is to optimize the parameters of the decision functions and to determine the leaf node distributions. To this end, a (possibly random sub-) set of training examples $\mathcal{S}$ is simultaneously pushed through the tree. Let us denote by $\mathcal{S}_i$ the training set reaching node i, where $\mathcal{S}_0 = \mathcal{S}$ at the root node with index 0. At each split node the incoming set $\mathcal{S}_i$ is divided into two disjunct, outgoing sets $\mathcal{S}_i^L$ and $\mathcal{S}_i^R$ which are sent to the left and right child nodes. The split is based on a decision function operating on the feature vectors of incoming training examples. Most commonly used split functions are so called axis-aligned functions $f_{\mathbf{v},\tau}$, defined as:

$$f_{\mathbf{v},\tau} \hat{=} (\mathbf{v}\cdot\mathbf{x} \geq \tau) \quad , \tag{2}$$

where $\mathbf{v}$ is a m-dimensional binary (random) vector and $\tau \in \mathbb{R}$ is a threshold. Note that $\mathbf{v}$ has only one non-zero entry and permits thereby to select one dimension from the m-dimensional feature space. τ is then either (randomly) drawn from the range of the feature values, or optimized via exhaustive search. Based on the decision function the training examples are separated into two subsets.

Following a greedy optimization strategy, different (randomly generated) split function candidates are evaluated and the most discriminative one is found based on maximizing an objective function such as the information gain:

$$I(\mathcal{S}_i, \mathcal{S}_i^L, \mathcal{S}_i^R) = H(\mathcal{S}_i) - \sum_{j\in\{L,R\}} \frac{|\mathcal{S}_i^j|}{|\mathcal{S}_i|} H(\mathcal{S}_i^j) \quad , \tag{3}$$

where $H(\cdot)$ is the entropy. In case of classification with a finite set of discrete labels $\mathcal{C}$, H is defined as the Shannon entropy

$$H(\mathcal{S}) = -\sum_{\mathbf{y}\in\mathcal{C}} p(\mathbf{y}|\mathbf{x}) \log p(\mathbf{y}|\mathbf{x}) \quad , \tag{4}$$

where $p(\cdot)$ is the empirical class distribution estimated from the training set $\mathcal{S}$. Good split functions should maximize the information gain which minimizes the uncertainty of the empirical distributions. When the tree growing process reaches a predefined depth, iterative splitting of the training data stops. The current node becomes a leaf where the empirical distribution over the incoming training examples is stored. The tree depth has an impact on the generalization of the tree as it directly influences the resolution of the partition of the feature space.

As a consequence of the objective function in Eq. (3), the training procedure yields leaf nodes with peaked class distributions. At test time, an unseen data

point should take the same path along the tree nodes as training examples with similar features. The empirical distribution over those training examples would then provide a good prediction for the test point.

After setting out the basics of decision forests in a classification scenario, next we discuss our main contribution: a joint classification-regression model employed within the same forest.

2.2 Joint Classification-Regression

Classification forests have been widely used in practice. In this paper we argue that in some applications their discriminative power can be improved by a simple yet elegant modification within the learning procedure. So far, the training of classification forests is only based on the ground truth class labels. The key idea of our approach is to explore also the spatial structure of objects. In fact, the same ground truth, *i.e.* label maps, contain information about the shapes of objects, and in multi-class problems, about relative positions and spatial arrangement (see Fig. 1 for an example). The integration of this rich information into the supervised learning can yield better predictions. To this end, we formulate a joint classification-regression approach where the training objective is to increase both class and spatial consistency. We introduce two prediction variables where $\mathbf{c} \in \mathcal{C}$ corresponds to a one-dimensional discrete classification output, and $\mathbf{r} \in \mathbb{R}^n$ is a n-dimensional continuous regression variable. The role of this variable is described in detail in Sec. 2.3. For now, let us assume it captures some continuous shape parameters. Given the same input variable $\mathbf{x}$ as before, our goal is now to learn the joint probability $p(\mathbf{c}, \mathbf{r}|\mathbf{x})$. Using the chain rule, we can rewrite this joint distribution as $p(\mathbf{c}, \mathbf{r}|\mathbf{x}) = p(\mathbf{r}|\mathbf{c}, \mathbf{x})\, p(\mathbf{c}|\mathbf{x})$. In order to learn this distribution within the framework of decision forests, we define the joint entropy as

$$\begin{aligned} H(\mathcal{S}) &= -\sum_{\mathbf{c}\in\mathcal{C}} \int_{\mathbf{r}\in\mathbb{R}^n} p(\mathbf{c}, \mathbf{r}|\mathbf{x}) \log p(\mathbf{c}, \mathbf{r}|\mathbf{x})\, d\mathbf{r} \\ &= \underbrace{-\sum_{\mathbf{c}\in\mathcal{C}} p(\mathbf{c}|\mathbf{x}) \log p(\mathbf{c}|\mathbf{x})}_{\text{Shannon Entropy: } H_{\mathbf{c}}} + \underbrace{\sum_{\mathbf{c}\in\mathcal{C}} p(\mathbf{c}|\mathbf{x}) \left(-\int_{\mathbf{r}\in\mathbb{R}^n} p(\mathbf{r}|\mathbf{c}, \mathbf{x}) \log p(\mathbf{r}|\mathbf{c}, \mathbf{x})\, d\mathbf{r} \right)}_{\text{Weighted Differential Entropy: } H_{\mathbf{r}|\mathbf{c}}} . \end{aligned} \tag{5}$$

During training, we maximize the same objective function as defined in Eq. (3), where now the entropy becomes $H(\mathcal{S}) = H_{\mathbf{c}}(\mathcal{S}) + H_{\mathbf{r}|\mathbf{c}}(\mathcal{S})$.

The two entropies $H_{\mathbf{c}}$ and $H_{\mathbf{r}|\mathbf{c}}$ may live within quite different ranges depending on the problem and its dimensionality, and one of them could easily overrule the other one during optimization. Hence, we propose the following normalization step

$$H(\mathcal{S}) = \frac{1}{2} \left(\frac{H_{\mathbf{c}}(\mathcal{S})}{H_{\mathbf{c}}(\mathcal{S}_0)} + \frac{H_{\mathbf{r}|\mathbf{c}}(\mathcal{S})}{H_{\mathbf{r}|\mathbf{c}}(\mathcal{S}_0)} \right) , \tag{6}$$

where each entropy is normalized w.r.t. the root node entropy. This normalization maps both initial entropies at the root node to one, and the information gain measures the relative improvement w.r.t. the inherent entropy of the training set.

2.3 Spatial Consistency via Distance Regression

In order to capture the spatial information contained in the label maps, we employ Euclidean signed distance maps (SDMs) as an implicit representation of shape. Assuming there are n different objects to be segmented, we can determine n distance maps per training image. Note that we treat the background as an extra class, so we have $|\mathcal{C}| = n+1$ number of classes, and no distance map is computed for the background class. Also note, that it is not necessary that all objects are present in all images. In practice, we can make use of indicator variables encoding the presence of an object which allows us to ignore missing data in the computation of statistics. For sake of simplicity, in the following we assume that all objects are present in all images.

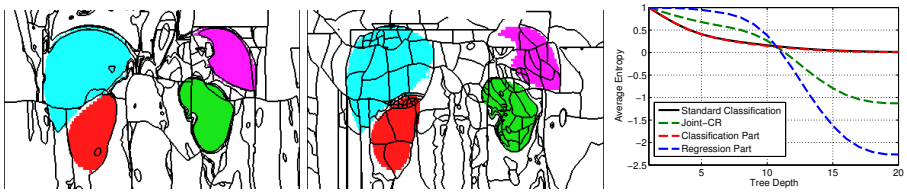
The distance maps allow us to assign n-dimensional vectors $\mathbf{r} = (d_1, ..., d_n)^\top$ to each pixel in the training set, where $d_\mathbf{c}$ is the distance of a pixel to the closest boundary point of the object with class index $\mathbf{c}$. Negative distances are assigned to pixels inside an object. This is an efficient way of enriching the training set to $\mathcal{S} = \{(\mathbf{x}_k, \mathbf{c}_k, \mathbf{r}_k)\}$, where now each data point carries both information about its class membership and its relative positions w.r.t. the shapes of all objects. The regression component $\mathbf{r}$ captures both shape and spatial layout of the objects, which in a common classification approach would remain hidden in the label maps. This supplementary information comes at no additional cost regarding annotations. This is a major advantage since acquiring ground truth data can be tedious and time-consuming, in particular, in the medical domain.

For efficient training of our joint model, we need a compact representation for the conditional distribution $p(\mathbf{r}|\mathbf{c}, \mathbf{x})$ which can be efficiently stored in the leaf nodes. We employ n-dimensional multivariate Normal distributions $p(\mathbf{r}|\mathbf{c}, \mathbf{x}) \hat{=} \mathcal{N}(\mu_{\mathbf{r}|\mathbf{c}}, \Sigma_{\mathbf{r}|\mathbf{c}}|\mathbf{r}, \mathbf{c}, \mathbf{x})$, one distribution per class label $\mathbf{c}$. Those can be efficiently stored by keeping only the means and covariance matrices. Additionally, Gaussian distributions have a closed-form definition for the differential entropy such that

$$H_{\mathbf{r}|\mathbf{c}} = \sum_{\mathbf{c}\in\mathcal{C}} p(\mathbf{c}|\mathbf{x}) \left(\frac{1}{2} \log \left[(2\pi e)^n |\Sigma_{\mathbf{r}|\mathbf{c}}| \right] \right) \quad , \tag{7}$$

where $|\cdot|$ denotes the determinant of a matrix.

Optimizing the information gain w.r.t. this entropy encourages splits which reduce the covariance over spatial location. This is the case when elements within subsets belonging to the same class are also spatially consistent. In fact, the regression component acts as a *learned* spatial regularization. In order to demonstrate this effect, we perform a small experiment. We take one 2D image (a coronal slice from a 3D CT scans) for training a single tree using the standard



(a) Classification (b) Joint CR (c) Entropy Progression

Fig. 2. (a,b) Leaf node region maps overlaid on ground truth segmentation. The maps illustrate the spatial regularization effect of the regression component. **(c)** Progression of different parts of the joint entropy (Eq. (6)) compared to standard classification.

classification objective function, and another tree using our joint objective function. To visualize the resulting "clustering" of training points, we use the same image at test time and store for each pixel the index of the reached leaf node. From these index maps we extract the cluster regions as shown in Fig. 2(a,b). Each closed region corresponds to a particular leaf node in the corresponding tree. At the same tree depth, training jointly on the combined classification-regression objective yields leaf nodes with clusters of training examples which are both consistent in terms of class membership *and* spatial location.

Robust Parameter Estimation The regression part of our joint predictor model requires estimation of means and covariances of the corresponding Gaussians $\mathcal{N}(\mu_{\mathbf{r}|\mathbf{c}}, \Sigma_{\mathbf{r}|\mathbf{c}}|\mathbf{r}, \mathbf{c}, \mathbf{x})$. This is commonly done via maximum likelihood (ML) estimation. Since we estimate the empirical distributions conditioned on the class label, the sample size for a particular distribution can become quite small. In order to overcome statistical problems when only few samples are available, we employ a more robust Bayesian estimation where the parent distribution of a child node plays the role of the prior. The mean is then estimated as

$$\mu_{\mathbf{r}|\mathbf{c}}^{\text{child}} = \frac{|\mathcal{S}_{\mathbf{r}|\mathbf{c}}^{\text{child}}|}{\kappa + |\mathcal{S}_{\mathbf{r}|\mathbf{c}}^{\text{child}}|}\, \bar{\mu}_{\mathbf{r}|\mathbf{c}}^{\text{child}} + \frac{\kappa}{\kappa + |\mathcal{S}_{\mathbf{r}|\mathbf{c}}^{\text{child}}|}\, \mu_{\mathbf{r}|\mathbf{c}}^{\text{parent}} \quad . \tag{8}$$

The covariance matrix is then computed as

$$\Sigma_{\mathbf{r}|\mathbf{c}}^{\text{child}} = \frac{|\mathcal{S}_{\mathbf{r}|\mathbf{c}}^{\text{child}}|}{Z}\, \bar{\Sigma}_{\mathbf{r}|\mathbf{c}}^{\text{child}} + \frac{\nu + n - 1}{Z}\, \Sigma_{\mathbf{r}|\mathbf{c}}^{\text{parent}} + \frac{\kappa\, |\mathcal{S}_{\mathbf{r}|\mathbf{c}}^{\text{child}}|}{Z\, (\kappa + |\mathcal{S}_{\mathbf{r}|\mathbf{c}}^{\text{child}}|)}\, \Psi_{\mathbf{r}|\mathbf{c}} \quad , \tag{9}$$

where $Z = \nu + n - 1 + |\mathcal{S}_{\mathbf{r}|\mathbf{c}}^{\text{child}}|$ and $\Psi_{\mathbf{r}|\mathbf{c}} = (\mu_{\mathbf{r}|\mathbf{c}}^{\text{parent}} - \bar{\mu}_{\mathbf{r}|\mathbf{c}}^{\text{child}})(\mu_{\mathbf{r}|\mathbf{c}}^{\text{parent}} - \bar{\mu}_{\mathbf{r}|\mathbf{c}}^{\text{child}})^{\top}$. Variables $\bar{\mu}_{\mathbf{r}|\mathbf{c}}^{\text{child}}$ and $\bar{\Sigma}_{\mathbf{r}|\mathbf{c}}^{\text{child}}$ are ML estimates of mean and covariance computed over the subset $\mathcal{S}_{\mathbf{r}|\mathbf{c}}^{\text{child}}$. Variables $\mu_{\mathbf{r}|\mathbf{c}}^{\text{parent}}$ and $\Sigma_{\mathbf{r}|\mathbf{c}}^{\text{parent}}$ correspond to the mean and covariance of the parent node. κ and ν are two parameters which permit to control the trade-off between the prior and the empirical information w.r.t. sample size. In fact, when the number of training examples $|\mathcal{S}_{\mathbf{r}|\mathbf{c}}^{\text{child}}|$ is sufficiently large ($|\mathcal{S}_{\mathbf{r}|\mathbf{c}}^{\text{child}}| >> \kappa, \nu$), the ML estimates dominate. When the number of training samples gets closer to the values of κ and ν the estimate of $\Sigma_{\mathbf{r}|\mathbf{c}}^{\text{child}}$ relies more on the parent.

2.4 Forest Predictions

Our joint classification-regression model allows to make two kinds of predictions at test time. The obvious one is regarding the most probable class label given a new data point, *i.e.* a pixel of a test image. This MAP estimate can be obtained by simply computing

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}\in\mathcal{C}} p(\mathbf{c}|\mathbf{x}) \quad . \tag{10}$$

Note, that test efficiency from a computational perspective is exactly the same as with standard classification forests. By obtaining the labels for all pixels, we determine the multi-object segmentation of the image.

We can also make predictions regarding the regression component. The most probable estimate of object distances for a pixel can be obtained by

$$\begin{aligned} \hat{\mathbf{r}} &= \arg\max_{\mathbf{r}} p(\mathbf{r}|\mathbf{x}) \\ &= \arg\max_{\mathbf{r}} \sum_{\mathbf{c}\in\mathcal{C}} p(\mathbf{r}|\mathbf{c},\mathbf{x}) p(\mathbf{c}|\mathbf{x}) \quad , \end{aligned} \tag{11}$$

which requires some sort of mode finding algorithm. Based on our Gaussian model, an alternative, robust estimate can be obtained via the mixture mean

$$\tilde{\mathbf{r}} = \sum_{\mathbf{c}\in\mathcal{C}} p(\mathbf{c}|\mathbf{x}) \mu_{\mathbf{r}|\mathbf{c}} \quad . \tag{12}$$

The regression allows us to estimate SDMs which could be of great use for instance in object alignment applications. One could think of defining a (weighted) matching criterion on both image intensities and regressed SDMs. The SDM part could potentially make the alignment less sensitive to initialization and more robust w.r.t. large transformations. The focus in this paper is on the segmentation part, and we are mainly interested in the label maps obtained via Eq. (10). However, we will also show results of SDM regression in the following section.

3 Experimental Validation

We evaluate our approach on the task of multi-organ segmentation in 3D medical CT scans. To this end, we collected a large dataset of 80 highly variable patient scans, in which 6 major organs have been manually delineated by an expert. The set of organs include liver, spleen, left and right kidney, left and right pelvic bone. To demonstrate the potential of our joint classification-regression strategy, we aim at isolating the effect of the proposed objective function, and therefore compare it directly with standard classifcation forests. The challenges in multi-organ segmentation arise from overlapping intensity profiles of different organs, variability in patient anatomy, presence of pathologies, and image noise. However, the human anatomy exhibits a highly structured spatial arrangement of inner parts. Hence, our approach is paricularly suitable for this task.

3.1 Experimental Setup and Training Parameters

For both methods, standard classification forest and our joint approach, we use the same fixed set of parameters. We train decision forests with 50 trees and a maximum tree depth of 20. We use bagging during training, which means each tree is trained on a random subset containing 10% of the total number of training image points. At each split node, we evaluate 100 different features from a pool of 1000 randomly generated features. For each feature, and the corresponding set of feature responses from the training points, we try 10 different thresholds uniformly distributed along the range of responses.

We employ five different types of features, where four of them are variants of 3D box features efficiently computed on integral images [21]: (i) a simple look-up of intensity in a smoothed version of the input image (Gaussian smoothing with $\sigma = 2$mm), (ii) average intensity in a randomly sized box centered at the image point, (iii) average intensity in a randomly sized box displaced by a random offset from the image point, (iv) intensity difference between the local intensity and a displaced box as in feature (iii), (v) intensity difference between two displaced boxes as defined in (iii). These features can capture both local and long-range contextual visual information. The range of the box sizes varies between 10 and 100mm. The displacements of boxes are drawn from an [0,100]mm interval. Concerning the Gaussian update for the mean and covariance estimation within the nodes, we choose $\kappa = 10$ and $\nu = 10$.

Fig. 2(c) shows the progress along tree depth of different parts of the entropy averaged over all trees. We make the following observations: i) the classification part $H_{\mathbf{c}}$ progresses almost identical compared to standard classification; ii) the regression part $H_{\mathbf{r}|\mathbf{c}}$ decreases mainly after a tree depth of 10.

3.2 Results

We split the 80 CT scans in two non-overlapping sets with each 40 scans and then perform a two-fold cross-validation. Hence, we can report overall segmentation errors computed on all 80 scans. The quantitative results for individual organs and the average performance are summarized in Fig. 3. Further qualitative results are shown in Fig. 4. We report errors w.r.t. ground truth annotations over four different segmentation scores, namely Dice's similarity coefficient (DSC) measuring the agreement between label maps (also known as F-score combining precision and recall into one value), and three surface distance measures. The mean surface distance (MSD), root-mean-square surface distance (RMS-SD), and Hausdorff distance (HD) are computed by determining the euclidean distances between segmentation boundaries extracted from the label maps. Note, that medical scans are always metrically calibrated (while the actual physical resolution between images varies). The unit of the last three errors is therefore in millimeters. All four scores indicate an improved performance when using our joint classification-regression approach. It is important to note, that both methods have access to exactly the same feature space. The difference in the segmentation results stems only from the modification of the training objective

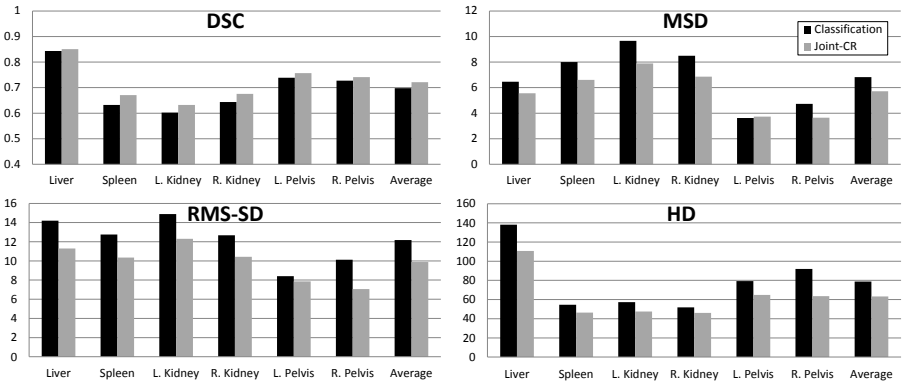


Fig. 3. Segmentation errors over four different scores. DSC measures the agreement between prediction and ground truth where 1 indicates perfect results. MSD, RMS-SD, and HD determine the surface distance in millimeters between prediction and ground truth where 0 indicates perfect results. Scores for classification forests are the black bars on the left, scores for our joint classification-regression are the gray bars on the right. All four scores indicate improved segmentation results for our approach.

function, which favors features in the greedy optimization which are yield both class and spatial consistency in the splits.

In particular, the improvement w.r.t. RMS-SD and HD is important. Both measures are sensitive to segmentation errors with larger distances. Here, the regularization effect of the regression component helps in removing outliers. This is confirmed by visual inspection of the qualitative results in Fig. 4. We observe that the segmentations for our joint approach are spatially more consistent and spurious results present in the standard classification are suppressed. We also show exemplary distance maps for the liver and left kidney. The regressed distance at each image point is the mixture mean as defined in Eq. (12).

4 Conclusion

We propose joint classification-regression forests as a novel supervised learning approach for the segmentation of spatially structured objects. Our experiments demonstrate that joint optimization yields superior results with both class and spatial consistency. This is achieved via a simple modification of the training objective combined with efficient representation of shape regression at no additional cost regarding annotations. A promising direction, where our method could be of direct use, is learning application-specific energy functions – *e.g.* in the context of random fields [12, 15]. Here, our joint model could be used to learn strong unaries which exhibit spatial smoothness learned from the training data. Other tasks, such as human pose estimation [5, 17] could also benefit from joint learning. In conclusion, we believe our model adds an important component to the framework of decision forests beyond the task of pixel-wise classification.

References

1. Breiman, L.: Random Forests. Machine Learning **45**(1) (2001) 5–32
2. Criminisi, A., Shotton, J., Konukoglu, E.: Decision Forests: A Unified Framework. Foundations and Trends in Computer Graphics and Vision **7**(2–3) (2011)
3. Ho, T.K.: Random Decision Forests. In: ICDAR. Volume 1. (1995) 278–282
4. Ho, T.K.: The Random Subspace Method for Constructing Decision Forests. PAMI **20**(8) (1998) 832–844
5. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-Time Human Pose Recognition in Parts from a Single Depth Image. In: CVPR. (2011) 1297–1304
6. Amit, Y., Geman, D.: Shape Quantization and Recognition with Randomized Trees. Neural Computation **9** (1997) 1545–1588
7. Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E.: Regression Forests for Efficient Anatomy Detection and Localization in CT Studies. In: MICCAI Workshop on Medical Computer Vision. (2010)
8. Bosch, A., Zisserman, A., Munoz, X.: Image Classification Using Random Forests and Ferns. In: ICCV. (2007)
9. Maree, R., Geurts, P., Piater, J., Wehenkel, L.: Random Subwindows for Robust Image Classification. In: CVPR. (2005)
10. Caruana, R., Karampatziakis, N., Yessenalina, A.: An Empirical Evaluation of Supervised Learning in High Dimensions. In: ICML. (2008) 96–103
11. Yin, P., Criminisi, A., Essa, I., Winn, J.: Tree-based Classifiers for Bilayer Video Segmentation. In: CVPR. (2007) 1–8
12. Payet, N., Todorovic, S.: $(RF)^2$ Random Forest Random Field. In: NIPS. (2010)
13. Kontschieder, P., Rota Buló, S., Bischof, H., Pelillo, M.: Structured class-labels in random forests for semantic image labelling. In: ICCV. (2011)
14. Montillo, A., Shotton, J., Winn, J.E., Iglesias, E., Metaxas, D., Criminisi, A.: Entangled Decision Forests and their Application for Semantic Segmentation of CT Images. In: IPMI. (2011) 184–196
15. Nowozin, S., Rother, C., Bagon, S., Sharp, T., Yao, B., Kohli, P.: Decision Tree Fields. In: ICCV. (2011)
16. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial Structures for Object Recognition. IJCV **61**(1) (2005) 55–79
17. Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A.: Efficient Regression of General-Activity Human Poses from Depth Images. In: ICCV. (2011) 415–422
18. Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V.: Hough Forests for Object Detection, Tracking, and Action Recognition. PAMI **33**(11) (2011) 2188–2202
19. Cootes, T., Edwards, G., Taylor, C.: Active Appearance Models. PAMI **23**(6) (2001) 681–685
20. Boykov, Y., Funka-Lea, G.: Graph Cuts and Efficient N-D Image Segmentation. IJCV **70**(2) (2006) 109–131
21. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. IJCV **57**(2) (2004) 137–154

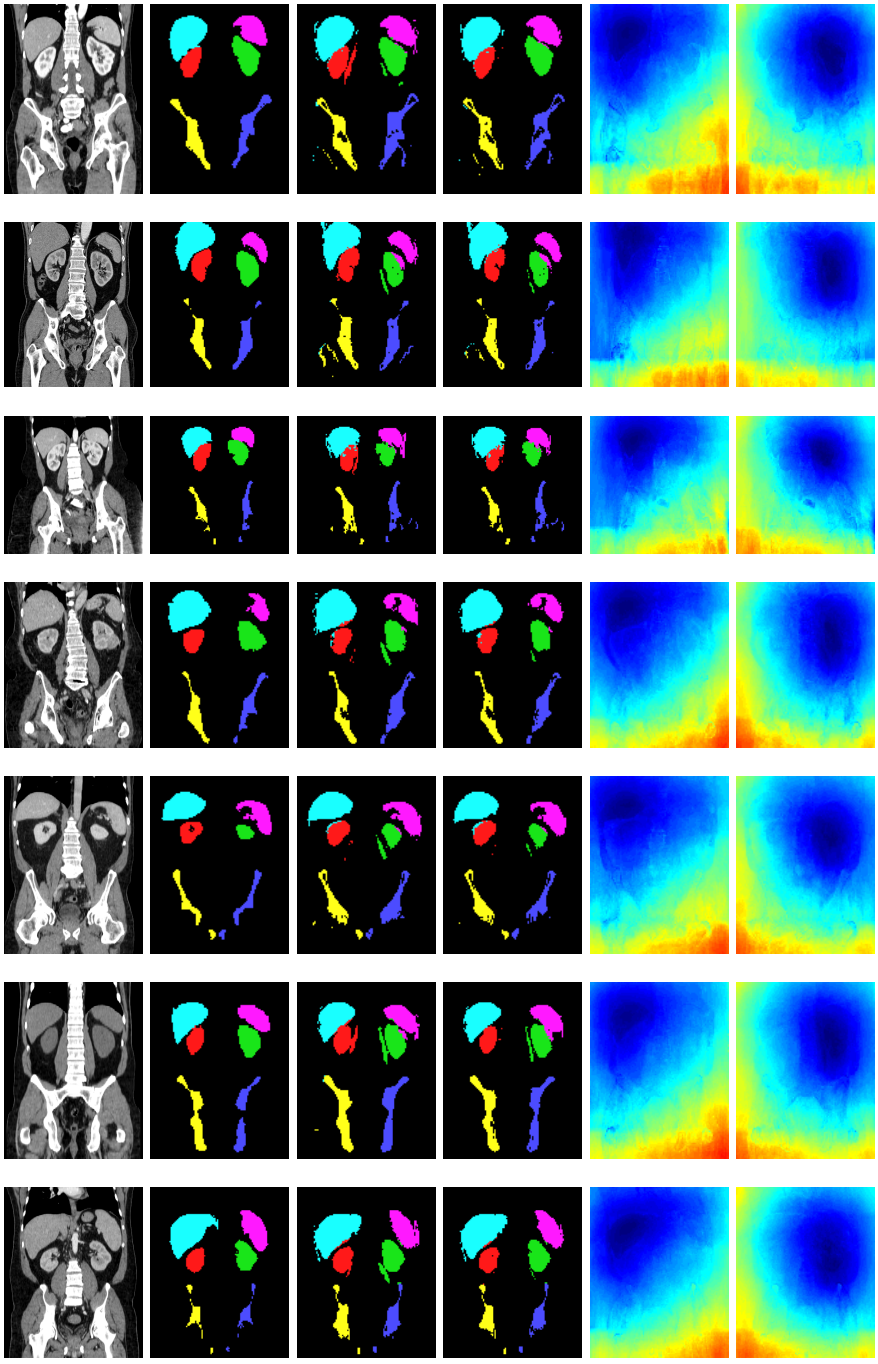


Fig. 4. From left to right: Slice from 3D input image, ground truth segmentation, MAP estimate of standard classification forest, MAP estimate of our joint approach, regressed distance maps for liver and left kidney obtained via Eq. (12).