Joint Supervoxel Classification Forest for Weakly-Supervised Organ Segmentation

Fahdi Kanavati[1], Kazunari Misawa[2], Michitaka Fujiwara[3], Kensaku Mori[4], Daniel Rueckert[1], and Ben Glocker[1]

[1] Biomedical Image Analysis Group, Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, UK
[2] Aichi Cancer Center, Nagoya 464-8681, Japan
[3] Nagoya University Hospital, Nagoya 466-0065, Japan
[4] Information and Communications, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

Abstract. This article presents an efficient method for weakly-supervised organ segmentation. It consists in over-segmenting the images into object-like supervoxels. A single joint forest classifier is then trained on all the images, where (a) the supervoxel indices are used as labels for the voxels, (b) a joint node optimisation is done using training samples from all the images, and (c) in each leaf node, a distinct posterior distribution is stored per image. The result is a forest with a shared structure that efficiently encodes all the images in the dataset. The forest can be applied once on a given source image to obtain supervoxel label predictions for its voxels from all the other target images in the dataset by simply looking up the target's distribution in the leaf nodes. The output is then regularised using majority voting within the boundaries of the source's supervoxels. This yields sparse correspondences on an over-segmentation-based level in an unsupervised, efficient, and robust manner. Weak annotations can then be propagated to other images, extending the labelled set and allowing an organ label classification forest to be trained. We demonstrate the effectiveness of our approach on a dataset of 150 abdominal CT images where, starting from a small set of 10 images with scribbles, we perform weakly-supervised image segmentation of the kidneys, liver and spleen. Promising results are obtained.

1 Introduction

Large datasets of medical images are increasingly becoming available; however, only a small subset of images tend to be fully labelled due to the time-consuming task of providing manual segmentations. This is one of the main hurdles for conducting large scale medical image analysis. Recently, a method [11, 3] using random classification forests to estimate correspondences between pairs of images at the level of compact supervoxels has been proposed. Given a pair of images, the method [11] consists in training a forest per image, applying it on the other image, and then extracting mutual correspondences. While the method is efficient if the application is restricted to a pair of images, it does not scale

well when applied to a large dataset of images: obtaining correspondences between images in a dataset of n images would entail training n distinct forests and testing $n(n-1)$ forests. In this paper we similarly use random forests to estimate correspondences; however, we make modifications to make it applicable to a large dataset: (a) we use object-sized supervoxels instead of small compact supervoxels, (b) we train *one* shared forest for *all* the images instead of one forest per image. We do this by allowing the trees to share the same structure by performing joint optimisation at the nodes. All the images are thus encoded by the same forest structure, where each leaf node stores a distinct supervoxel label distribution per image. This method makes it possible to compute correspondences efficiently between all the image in a large dataset on a supervoxel-level by a simple look up process in the leaf nodes. The obtained correspondences can then be used to propagate semantic labels. We investigate using the proposed method in a weakly-supervised medical image segmentation setting on an abdominal CT dataset consisting of 150 images, starting only from 10 weakly-labelled images.

Related Work: In the computer vision community, there has been considerable research done in the domain of weakly supervised image segmentation [10, 6, 2], segmentation propagation [13, 7], and co-segmentation [17], where the minimal assumption is that a common object is present in all of the images. Other form of weak annotations could be included such as bounding boxes, scribbles or tags indicating the presence of some object of interest. The main motivation behind these methods is the idea that a large dataset containing similar objects is bound to have repeating patterns and shapes, so they could potentially be exploited to discover and jointly segment the common objects. One issue when attempting to apply some of the methods to medical images is scalability, where the main bottleneck is obtaining correspondences between the images. Some state-of-the-art methods [13] rely on dense pixel-wise correspondences, which is infeasible to apply to a large dataset of 3D medical images. In an attempt to overcome such issues, other methods advocate using superpixels in an image as a building block in unsupervised and weakly supervised segmentation [14, 18, 7], where feature descriptors are typically computed on a pixel-level and then aggregated within superpixels; however, descriptor choice is non-trivial and can still be computationally costly for 3D images depending on the type of features.

To investigate whether weak supervised segmentation can be performed efficiently on a large 3D medical dataset, we make use of random forests, which are one of the most popular supervised machine learning algorithms that have been used in medical image analysis [5, 12, 4, 20]. Their popularity comes from their flexibility, efficiency, and scalability. The random forest framework makes it possible to do feature selection from a large pool of features. Additionally, random forest can scale up efficiently to large data, like 3D images, especially when simple cuboid feature, coupled with integral images, are used. Due to the nature of medical images, where anatomical structures follow certain patterns, context around voxels plays a role in improving classification [19]. Random forest with offset cuboid features can efficiently exploit context in an image to improve the prediction accuracy.

2 Method

The problem is similar to one in [11], except that instead of pairs of images, we extend it to multiple images. Given a set of images $\mathcal{I} = \{I_i\}_1^N$ and their set of associated supervoxels $\mathcal{R} = \{R_{ik}\}$, $k = 1 \dots |C^i|$, the aim is to establish correspondences between supervoxels across all images. We note $C^i = 1, \dots, |C^i|$ as the index set of the supervoxels of image I_i.

2.1 Object-sized Supervoxels

We over-segment each image into a set of supervoxels using a 3D extension of the efficient graph-based segmentation algorithm [8]; it takes in three parameters: k, min_size, and σ (Gaussian smoothing). For more details about the parameters, please refer to [8]. The number of supervoxels generated depends on the image. This algorithm allows obtaining segments that potentially represent different anatomical structures and organs; it has been used as a component of region proposal algorithms [16]. However, it is not possible to accurately segment each organ such that each organ is contained in one supervoxel. Over-segmentation and under-segmentation still occurs. Given two neighbouring voxels p ad q, we modify the weighting term used in the algorithm to

$$w(p,q) = \frac{|I(p) - I(q)|}{\max(10, S_\beta(p))} \times \sqrt{G(p) \times G(q)}, \tag{1}$$

where $G(p)$ is the gradient magnitude obtained using the Sobel filter and helps in reducing the sensitivity to noise. $S_\beta(p)$ is a Gaussian filtered image, with smoothing parameter β, of the average of the absolute value of the gradient in the 3 directions; this term provides adaptive contrast normalisation. The proposed modification results in better looking supervoxels and reduces the impact of noise in CT images. In addition, during the post-processing step, we add an additional constraint where we only merge a pair of supervoxels if any of their sizes is less than min_size and their edge weight is less than m. Any remaining supervoxels that are smaller than min_size are excluded from the final output. Examples are shown in the Fig. 1 as well as in the supplementary material.

2.2 Joint Supervoxel Random Classification Forest

Due to lack of space, we directly describe the random forest framework as applied to our problem. A more extensive overview can be found at [4].

We train a single forest on all images simultaneously such that all the images are encoded in that single forest. Fig. 1 shows an overview of the proposed method, where only one tree is shown.

Node Optimisation: At each node, we perform joint optimisation of the feature selection, where the feature selected at each node is the one that maximises the sum of the total information gain from all the images.

$$IG(S) = \sum_i IG(S^i) \tag{2}$$

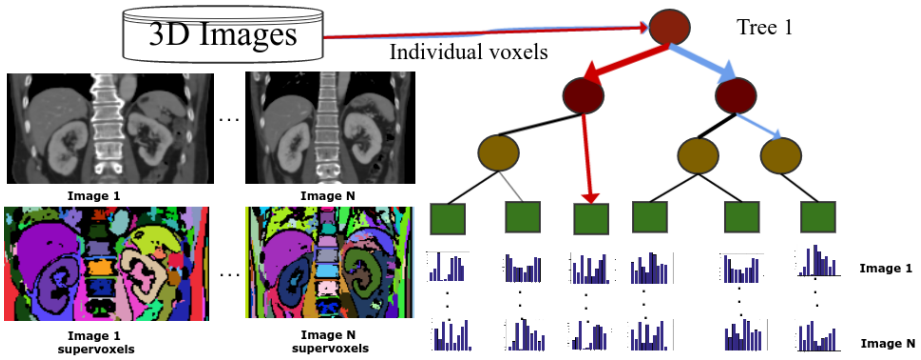

Fig. 1. An overview of the joint supervoxel forest. For illustration, only one tree is shown. Voxels from all the images in the dataset are used to train each tree. Each voxel has its supervoxel index as a label. At the nodes, we perform joint optimisation by maximising the sum of the information gain from all the images. At the leaf nodes, we store a distribution per image of the supervoxel indices, such that there could be up to N distributions stored at each leaf node. During testing, a voxel starts at the root node, and depending on its response to the binary split function at each node (circle), it is sent left or right until it reaches a leaf node (square). Once a voxel from a given image reaches a leaf node, it is possible to look up simultaneously its correspondences to all the other images.

where $IG(S^i)$ is the information gain of the ith image with samples S^i, and

$$IG(S^i) = H(S^i) - \sum_{j=\{L,R\}} \frac{|S^i_j|}{|S^i|} H(S^i_j) \ , \tag{3}$$

where $H(S) = -\sum_{c \in C} p(c) \log p(c)$ is the Shannon entropy and $p(c)$ is the normalised empirical histogram of the labels of the training samples in S^i.

In addition, we randomly select, with ratio r_s, at each node, a subset of the images to perform node optimisation on. This allows speeding up the training process and increasing the randomisation between the trees.

Leaf Nodes: At each leaf node, we store the posterior distributions $p_i(c = l|\mathbf{v}), l \in C^i, i = 1 \dots n$ of the supervoxel labels for each one of the n images. After training is done, the distributions for the *ith* image in the leaf nodes are computed by re-passing down *all* the image's voxels. We have found that this increases the accuracy of the correspondences, as the initial low sampling rate used during training is not enough to create an accurate distribution.

Appearance Features: Similarly to other methods, we use a set of context appearance features as they have been found to be quite effective and efficient [4, 9, 20] The features consist of local mean intensities and mean intensity differences between two different cuboid regions at different offsets, which are efficient to compute via the use of 3D integral images. The feature generation function $\psi(\mathbf{x}) : \mathbb{R}^3 \rightarrow \mathbb{R}$ takes as input the position $\mathbf{x}$ of the voxel and computes a feature value based on: a pair of offsets $(\Delta\mathbf{x}_0, \Delta\mathbf{x}_1) \in \mathbb{R}^3 \times \mathbb{R}^3$; a pair of size parameters $(\mathbf{s}_0, \mathbf{s}_1) \in \mathbb{R}^3 \times \mathbb{R}^3$, where a given $\mathbf{s}$ characterises the dimensions of a

cuboid centred at position $\mathbf{u}$; $B_s(\mathbf{u})$ is the mean intensity of the voxels within the cuboid centred at $\mathbf{u}$ and of size $\mathbf{s}$; and $b \in \{0, 1\}$ is a binary value that indicates whether to take the intensity difference between two cuboids or only the value from a single cuboid. Let $\kappa = \{\mathbf{s}_0, \mathbf{s}_1, \Delta\mathbf{x}_0, \Delta\mathbf{x}_1, b\}$ denote the set of parameters. Given some choice of values for κ, the feature response for a voxel at $\mathbf{x}$ in image I is: $\psi_\kappa(\mathbf{x}) = B_{s_0}(\mathbf{x} + \Delta\mathbf{x}_0) + b \times B_{s_1}(\mathbf{x} + \Delta\mathbf{x}_1)$.

Once the feature response $\psi_\kappa(\mathbf{x})$ has been evaluated for all samples at a given node m, the optimal value for the threshold τ_m is obtained via a grid search.

Establishing Correspondences: Once the forest has been trained, correspondences between all the n images in the dataset can be obtained by applying the forest once on each image. Once a voxel $\mathbf{v}_i$ from image I_i reaches a leaf node of a tree t, it gets assigned a set of probabilities $\{p_j^t(c = l|\mathbf{v}_i) \mid l \in C^j\}_{j=1\ldots n}$ for each one of the other images I_j in the dataset.

Correspondences on the supervoxel level are then obtained via majority voting, by aggregating all the probabilities of a supervoxel's voxels from all the trees and finding the labels that have maximum probabilities. So, given a supervoxel sv_k^i from image I_i, it gets assigned n labels $c_k^{ij}, j = 1 \ldots n$ obtained as follows:

$$c_k^{ij} = \arg\max_{c \in C^j} \sum_{t=1\ldots T} \sum_{v \in sv_k^i} p_j^t(c|\mathbf{v}), \tag{4}$$

Mutual Correspondences: The correspondences are pruned such that a match $A \rightarrow B$ is considered valid only if there is also a match from $B \rightarrow A$. This helps in pruning out false correspondences. For each supervoxel A, we get a correspondence set M_A consisting of all the supervoxels from the other images that match with it. We expand the set of correspondences to neighbouring ones such that if we have a mutual correspondence between $A \leftrightarrow B$, and $B \leftrightarrow C$, then a correspondence between $A \leftrightarrow C$ is created. In addition, in each set M_A, we prune out the elements that do not mutually match with at least 50% of the other elements in M_A.

2.3 Weakly-supervised Segmentation

Given a dataset where a subset has scribbles on organs, each object-sized supervoxel gets assigned the label of its organ scribble. We then use the obtained correspondences to propagate the organ labels, resulting in a larger subset of labelled images. Some images remain unlabelled, however, if they did not receive any correspondences. We therefore train another set of forests, one per organ; however, this time using the organs as labels and using the subset of images that have obtained a given organ label. Applying each organ label forest on *all* the images results in a probabilistic output as to whether a voxel is of a given organ or background. It is applied on all the images so as to potentially correct the initial over-segmentation errors and to segment the remaining unlabelled images. The probabilistic outputs (only probabilities larger than 0.85 are used) are fed in as the unary cost to graph-cut [1] so as to obtain regularised binary segmentation outputs. For each image, we then fuse the binary organ segmentations into a single image. The result is a fully-segmented dataset.

3 Experiments and Results

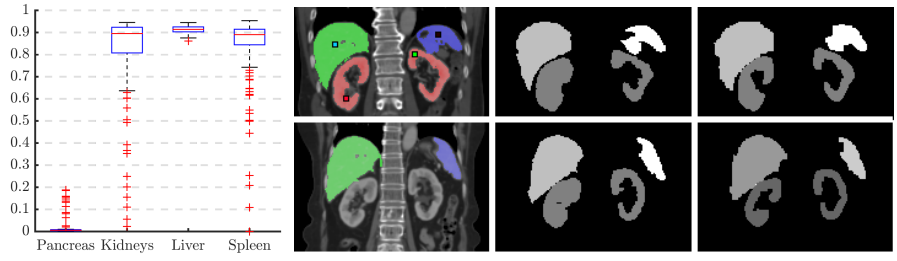

Fig. 2. Left: Final dice overlap score computed for the 150 images in the dataset at a $2mm$ spacing. We see that apart from a few outliers cases for the kidneys, liver and spleen, we get an overall median dice of 0.9 for all three organs. Poor results were obtained for the pancreas. Right: Image on the top left had organ scribbles (square dots), which were then assigned to its supervoxels (coloured overlay correspond to the supervoxels that were assigned an organ label. Full supervoxels are shown in Fig. 1). The bottom left image was unlabelled and did not have any organ scribbles; however, it received a liver and spleen label for its supervoxels via the the joint forest correspondences. The right column shows the segmentation output after applying the organ forest and graph-cut. We see that the bottom image gets segmentations for the kidneys.

Dataset: We use an abdominal CT dataset consisting of 150 distinct subjects, which mostly contains pathological cases. The 3D scans have an in-plane resolution of 512×512; the number of slices is between 238 and 1061. Voxel sizes vary from 0.55 to 0.82; slice spacing ranges from 0.4 to 0.8 mm. Manual organ segmentations of the liver, spleen, kidneys, and pancreas are available. **Experimental Set-up:** We generate dot scribbles on 10 randomly selected images, where each image receives 4 dot scribbles, representing the 4 organs. We repeat the experiment 10 times. Supervoxels are computed on images re-sampled to $2mm$, with $k = 80$, minimum size $3000mm^3$, $\sigma = 0.5mm$, $\beta = 5mm$, and $m = 50$ (all set empirically). For the joint forest, images are re-sampled to $6mm$; we train 20 trees, with a max depth of 14, max offset (in native resolution) of 200mm, cuboid sizes up to 32mm, 15 features/node, a grid search with 10 bins, min number of samples 5, $r_s = 50\%$ and a sampling rate of 1%. Node growing during training stops when the maximum depth is reached, the number of samples is less than 5, or there is no entropy improvement. For the organ forest, we use similar parameters, except that we increase the re-sampling to $3mm$, the features/node to 100, and the sampling rate to 5%. Graph-cut is applied using the pairwise Potts model on images of $2mm$ spacing, with $\lambda = 4$.

Results: From an initial set of 10 weakly organ labelled images (with dot scribbles), the number of images that received an organ supervoxel label via the correspondences were on average: 101, 110, 144, and 121 for the right kidney, left kidney, liver, and spleen, respectively. We report the dice overlap of the final

segmentation output with graph-cut applied on all the 150 images. For each image, the dice overlap was averaged from the 10 random runs; box plots of those averages are reported in Fig. 2, where we see that apart from the pancreas and a few outlier cases for the kidneys, liver, and spleen, we get good segmentation accuracy with extremely minimal user input. The reader is advised to refer to the supplementary material for more visual results from each step of the pipeline.

4 Discussion and Conclusion

In this paper, we have presented an efficient method for weakly-supervised organ segmentation. The method consisted in over-segmenting the images into object-sized supervoxels and training a single shared forest using all the images via joint node optimisation. The joint forest was then used to efficiently estimate correspondences between supervoxels in an abdominal CT dataset. These correspondences were used to propagate weak organ labels from a small set of 10 images to 140 unlabelled images. A second forest was then trained using organ labels on a supervoxel level. The organ forest was applied on all the images so as to correct potential inaccuracies in the over-segmentation and to segment any remaining unlabelled image. The probabilistic output from the forest was then used as the unary cost for graph-cut to regularise the output. Apart from poor results for the pancreas, which is difficult to segment due to its extremely deformable nature, the liver, kidney and spleen obtain good segmentation results, excluding a few outlier cases (a fully-supervised method [15] applied on the same dataset obtains a dice overlap of 94.9%, 93.6%, and 92.5% for liver, kidneys, and spleen, respectively.). Our method could potentially be used to provide coarse segmentation or mine a large dataset of medical images, with extremely minimal user input. The advantage of our method is that it is efficient, as images can be greatly down-sampled and the random forest framework is easily parallelisable. One limitation is that method is not able to handle extremely deformable organs, such as the pancreas. Another limitation is that the initial supervoxel over-segmentation parameters were determined empirically; however, the availability of a small subset of fully-labelled images could help in determining the parameters. An alternative would be generating multiple over-segmentations per image, as advocated by some methods [14, 7] (e.g. by using object proposals). We could then train a joint forest per over-segmentation set. The output from the multiple joint forests could then be used as an ensemble. Future work would involve investigating this, as well as investigating the use of alternative supervised algorithm (e.g. convolutional network), alternative image features to attempt to segment the pancreas and a more integrated interactive user input, which could help in correcting outliers and speeding up the interactive segmentation process.

References

1. Boykov, Y.Y., Jolly, M.M.P.: Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. ICCV 2001. 1, 105–112 vol.1 (2001)

2. Chen, X., Shrivastava, A., Gupta, A.: Enriching visual knowledge bases via object discovery and segmentation. In: Proceedings of the IEEE conference on CVPR. pp. 2027–2034 (2014)
3. Conze, P.H., Tilquin, F., Noblet, V., Rousseau, F., Heitz, F., Pessaux, P.: Hierarchical multi-scale supervoxel matching using random forests for automatic semi-dense abdominal image registration. In: IEEE ISBI (2017)
4. Criminisi, A., Shotton, J.: Decision forests for computer vision and medical image analysis. Springer Science & Business Media (2013)
5. Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E.: Regression forests for efficient anatomy detection and localization in ct studies. In: Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging. pp. 106–117. Springer (2011)
6. Deselaers, T., Alexe, B., Ferrari, V.: Weakly supervised localization and learning with generic knowledge. IJCV 100(3), 275–293 (2012)
7. Dutt Jain, S., Grauman, K.: Active image segmentation propagation. In: Proceedings of the IEEE Conference on CVPR. pp. 2864–2873 (2016)
8. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. IJCV (2004)
9. Glocker, B., Zikic, D., Haynor, D.R.: Robust registration of longitudinal spine ct. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014, pp. 251–258. Springer (2014)
10. Grauman, K., Darrell, T.: Unsupervised learning of categories from sets of partially matching image features. In: CVPR, 2006 IEEE Computer Society Conference on. vol. 1, pp. 19–25. IEEE (2006)
11. Kanavati, F., Tong, T., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D., Glocker, B.: Supervoxel classification forests for estimating pairwise image correspondences. Pattern Recognition 63, 561–569 (2017)
12. Montillo, A., Shotton, J., Winn, J., Iglesias, J.E., Metaxas, D., Criminisi, A.: Entangled decision forests and their application for semantic segmentation of ct images pp. 184–196 (2011)
13. Rubinstein, M., Liu, C., Freeman, W.T.: Joint inference in weakly-annotated image datasets via dense correspondence. IJCV 119(1), 23–45 (2016)
14. Russell, B.C., Efros, A.A., Sivic, J., Freeman, W.T., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. Proceedings of the IEEE Computer Society Conference on CVPR 2, 1605–1612 (2006)
15. Tong, T., Wolz, R., Wang, Z., Gao, Q., Misawa, K., Fujiwara, M., Mori, K., Hajnal, J.V., Rueckert, D.: Discriminative dictionary learning for abdominal multi-organ segmentation. Medical image analysis 23(1), 92–104 (2015)
16. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. IJCV 104(2), 154–171 (2013)
17. Vicente, S., Rother, C., Kolmogorov, V.: Object cosegmentation. In: CVPR, 2011 IEEE Conference on. pp. 2217–2224. IEEE (2011)
18. Xu, J., Schwing, A.G., Urtasun, R.: Learning to segment under various forms of weak supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3781–3790 (2015)
19. Zhou, S.: Chapter 1 - introduction to medical image recognition, segmentation, and parsing. In: Zhou, S.K. (ed.) Medical Image Recognition, Segmentation and Parsing, pp. 1 – 21. Academic Press (2016)
20. Zikic, D., Glocker, B., Criminisi, A.: Encoding atlases by randomized classification forests for efficient multi-atlas label propagation. Medical image analysis (Jul 2014)