Supervoxel Classification Forests for Estimating Pairwise Image Correspondences

Fahdi Kanavati[1], Tong Tong[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 paper proposes a general method for establishing pairwise correspondences, which is a fundamental problem in image analysis. The method consists of over-segmenting a pair of images into supervoxels. A forest classifier is then trained on one of the images, the source, by using supervoxel indices as voxelwise class labels. Applying the forest on the other image, the target, yields a supervoxel labelling which is then regularized using majority voting within the boundaries of the target's supervoxels. This yields semi-dense correspondences in a fully automatic, efficient and robust manner. The advantage of our approach is that no prior information or manual annotations are required, making it suitable as a general initialisation component for various medical imaging tasks that require coarse correspondences, such as, atlas/patch-based segmentation, registration, and atlas construction. Our approach is evaluated on a set of 150 abdominal CT images. In this dataset we use manual organ segmentations for quantitative evaluation. In particular, the quality of the correspondences is determined in a label propagation setting. Comparison to other state-of-the-art methods demonstrate the potential of supervoxel classification forests for estimating image correspondences.

1 Introduction

Establishing correspondences between images is a fundamental and important problem in many medical image analysis tasks. To this end, dedicated image registration techniques have been developed and successfully employed in fully automated analysis pipelines [15]. Many of these techniques work best when applied on particular types of images, such as brain scans, where simple initialisation strategies work well. In general settings, however, the images to be registered might capture very different fields of view, as it is often the case in pre- and post-operative abdominal scans. In such settings, establishing an initial alignment can be quite challenging if no prior information is available. It can be beneficial to utilize anatomy recognition and landmark detection methods which

provide spatial priors for registration [7]. However, this requires an annotated image database for training. Obtaining a large number of manually annotated images can be tedious, costly and time-consuming.

Contribution: We propose a general method for establishing initial pairwise correspondences which does not require any prior information or manual annotations. We employ classification forests [2], but in contrast to previous work class labels for training are generated automatically. Our method consists of over-segmenting a pair of images into supervoxels. We then train a forest classifier on one of the images – the source image – by using its supervoxels indices as voxelwise class labels. Applying the forest on the other image – the target image – yields a supervoxel label prediction for each of its voxels. Majority voting is then carried out within the supervoxels of the target image where each voxel casts a vote as to what the final supervoxel label should be. The final labelling yields correspondences between the supervoxels of the two images. Supervoxels are an ideal representation for semi-densely distributed correspondences relaxing the one-to-one matching assumption between images. Establishing supervoxel correspondences between two images solves the initialization problem for many image analysis tasks such as atlas/patch-based segmentation [8, 4], registration, and atlas construction.

Related Work: Random forests [2], as a supervised machine learning technique, have found many successful applications in medical image analysis [6, 10, 5, 14]; this is mainly due to their accuracy, robustness, and scalability. They rely on the availability of labelled images which is contrast to the approach taken here where labels are generated automatically. While traditionally, forests are trained on a database containing many images, recently, the idea of encoding a single labelled image (or "atlas") as a forest [14] has been proposed in the context of multi-atlas label propagation. This has inspired our idea of using the atlas-forest approach for learning image correspondences from a single source image, which is labelled automatically via a supervoxelisation. Supervoxels – and their 2D counterpart, superpixels – have found many applications in computer vision [12, 9]. They allow the grouping of voxels into locally consistent regions that have similar properties thereby reducing redundancy and computational complexity. Supervoxels are mainly used within segmentation pipelines. We are not aware of previous work that has used supervoxels as label entities in classification forests, in particular, with the aim of establishing image correspondences.

2 Methods

2.1 Problem Formulation

The aim of our method is to estimate correspondences between a set of image regions, i.e. supervoxels. Let I_i be an image that is over-segmented into an indexed set $\mathcal{SV}^i = (sv_k^i)_{k \in C^i}$ of distinct supervoxels sv_k^i. The image therefore consists of $|\mathcal{SV}^i|$ supervoxels with the index set $C^i = \{1, ..., |\mathcal{SV}^i|\}$ denoting the distinct labels of the supervoxels. Each supervoxel $sv_k^i = \{\mathbf{v}_l^i\}_1^{|sv_k^i|}$ in turn is a

set of voxels $\mathbf{v}_l^i$. With N^i representing the total number of voxels in the image, we would have $\sum_k |sv_k^i| = N^i$.

Establishing correspondences from an image I_i to an image I_j consists of finding a mapping function g^i that maps each supervoxel $sv_k^j \in \mathcal{SV}^j$ to a value/label in the index set C^i so that $\forall k \in C^j, \exists c \in C^i \mid g^i(sv_k^j) = c$. We propose to use random classification forests to learn the mapping function g^i.

2.2 Random Forests

First we give a brief overview of random forests [2] when applied to a single 3D image. An excellent in depth review can be found at [5]. Random forests are a collection of binary decision trees. They involve two stages: training and testing. The data used for training the forest consists of all the voxels from a single image. We denote the training set as $\mathcal{S} = \{\mathbf{v}_k, c_k\}_1^N$ with $c_k \in C$ being the label of voxel $\mathbf{v}_k$. A tree consists of a set of nodes such that each node can either be a leaf node or has two child nodes.

Each m^{th} node has a binary weak classifier $f(\mathbf{v}, \theta) = [\phi_m(\mathbf{v}) - \tau_m]$ with $\theta_m = \{\phi_m, \tau_m\}$; $\phi_m(\mathbf{v})$ is an appearance feature and τ_m is a threshold. The weak classifier serves as a split function that determines whether a given sample should go down the left or the right child node. For a set of samples S^m arriving at the m^{th} node, different θ_m values yield different disjoint subsets S_L^m and S_R^m.

Training a tree involves finding at each m^{th} node the optimal $\hat{\theta}_m$, via maximisation of an objective function $h(S^m, S_L^m, S_R^m, \theta^m)$. Each node has access to a limited number n_f of randomly generated values for θ^m. The randomness ensures that each tree ends up being unique. Starting from the root node, the samples are recursively split up into two subsets based on the optimal split, with a subset going down each child node. Once a stopping criteria is met – such as maximal depth or minimal sample count – the node becomes a leaf and the distribution of samples that reached it is stored as a posterior probability $p_t(c|\mathbf{v})$.

When *testing* on a new target image, its voxels are passed down the learnt tree going left or right, depending on their response to the split function found during training, until reaching a leaf node (Fig. 1). The outputs from all the T trees in the forest are then combined by averaging: $p(c|\mathbf{v}) = \frac{1}{T}\sum_t p_t(c|v)$. The final voxel label is obtained by selecting the maximum $\hat{c} = \arg\max_c p(c|\mathbf{v})$.

For a *classification forest*, the objective function h at a node with samples S is the information gain $H(S) - \sum_{i=\{L,R\}} \frac{|S_i|}{|S|} H(S_i)$, where $H(S)$ is the Shannon entropy $-\sum_{c\in C} p(c)\log p(c)$ and $p(c)$ is the normalised empirical histogram of the labels of the training samples in S.

A *regression forest*, unlike a classification forest, maps an input into a continuous output. To use a regression forest to output correspondences, each voxel $\mathbf{v}$ has its position assigned as its label $\mathbf{c} = (x, y, z)$. The objective function used in this case [3] is the error of the fit $\sum_{\mathbf{v}\in S} (\mathbf{c} - \bar{\mathbf{c}})^2 - \sum_{i=\{L,R\}} \sum_{\mathbf{v}\in S_i} (\mathbf{c} - \bar{\mathbf{c}})^2$ with $\bar{\mathbf{c}}$ being the mean position vector for all the points at a node with samples S.

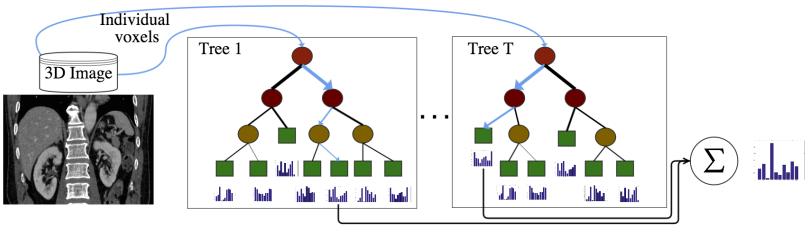

Fig. 1. An overview of random forests. All voxels of a single image are fully used to train each tree. During testing, a voxel starts at the root node and depending on its response to the binary split function at each node (circle), it is sent left or right until it reaches a leaf node (square). The posterior probability distributions from the reached leaf nodes are then averaged to obtain a final label posterior distribution.

Appearance Features Similarly to [14], we use a set of context appearance features with offsets up to 200mm; large offsets have been found useful [6] for discriminating between organs. The features used consist of intensities and differences between intensities in two different regions. The feature function $\phi(\mathbf{v})$ mentioned in Sec. 2.2 is characterised by: an offset $\Delta\mathbf{x} \in \mathbb{R}^3$ and a 3D box $B_s(\mathbf{x})$ centred at $\mathbf{x}$ with a size parameter $s \in \mathbb{R}^3$. For a voxel with $\mathbf{v} \in \mathbb{R}^3$ representing its position, $\phi(\mathbf{v})$ can be any of the following:

1. Mean intensity of local box: $\langle I(B_s(\mathbf{v}))\rangle$
2. Difference of intensity of local point and mean intensity of offset box: $I(\mathbf{v}) - \langle I(B_s(\mathbf{v}+\Delta\mathbf{x}))\rangle$
3. Difference of mean intensity of local box and mean intensity of offset box: $\langle I(B_s(\mathbf{v}))\rangle - \langle I(B_s(\mathbf{v}+\Delta\mathbf{x}))\rangle$
4. Difference of a pair of offset box means $\langle I(B_s(\mathbf{v}_1+\Delta\mathbf{x}_1))\rangle - \langle I(B_r(\mathbf{v}_2+\Delta\mathbf{x}_2))\rangle$

Once the response $\phi(\mathbf{v})$ has been evaluated for all samples at a given node, the optimal value for the threshold τ is obtained by uniformly dividing the response space into $n_{\text{thresholds}}$ and choosing the value the maximises the information gain.

2.3 Supervoxel Classification Forest (SVF)

In our proposed method we encode a single image into a classification forest as in [14]; however, instead of using organ labels, the label of each voxel is the index of the *supervoxel it belongs to*. Random forests can easily handle a large number of labels making them suitable for this task.

To generate supervoxels, we use the efficient SLIC superpixel [1] algorithm. It performs k-means clustering using intensities balanced with the euclidean distance as a distance measure; it takes as input the size of the desired supervoxels and a compactness parameter that enforces regularity in the supervoxel shape. The output is a set of approximately regularly spaced supervoxels that tend to follow intensity boundaries.

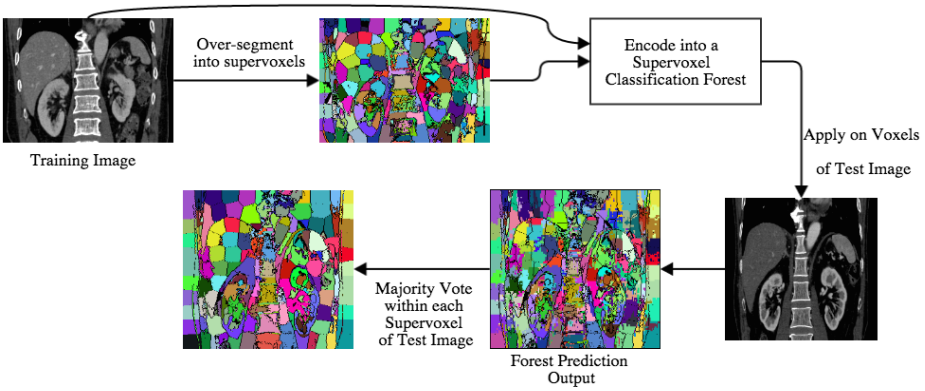

Fig. 2. : Proposed method for establishing correspondences at a supervoxel level. First, the training image is segmented into supervoxels (randomly coloured) which are then used as labels to train a classification forest using all the voxels in the image. Applying the forest on a test image yields a supervoxel label prediction for each voxel that does not necessarily follow the test image's supervoxel boundaries. As a final step, the voxels within each supervoxel in the test image cast votes as to what its label should be. Same colour indicates a match between supervoxels in the training and the test image.

Given the training image I_i and its set of labels $|C^i|$ (or in this case supervoxel indices), a training set is constructed using all the voxels in the image $\{\mathbf{v}_k^i, c_k\}_1^N$ with $c_k \in C^i$ and we use it to train an SVF as described in Sec. 2.2.

When applying the forest on a test image, the label predictions from the forest tend to be noisy (Fig. 2); therefore, we perform as a final step a majority voting within each supervoxel of the test image based on the predicted labels of their voxels. Each supervoxel sv_k^j in the test image I_j receives votes from each one of its voxels as to what its label from C^i should be. The final supervoxel label of sv_k^j is obtained by selecting the label with the maximum votes $c_k^j = \arg\max_{c \in C^i} \sum_{\mathbf{v} \in sv_k^j} p(c|\mathbf{v})$.

3 Experiments and Results

Ground truth data for one-to-one correspondences between images is hard to obtain. Therefore, to quantitatively evaluate our method, we test it in a simple multi-atlas label propagation (MALP) setting. We do this as MALP is an application that inherently requires establishing correspondences between images in order to propagate labels. Most state-of-the-art methods in MALP such as in [13, 11] use affine registration as a first step to give an initial set of dense correspondences between the atlases and the target image before proceeding with a more sophisticated label propagation scheme. Although affine registration is less accurate than doing non-rigid registration, it is used because it is more efficient. As random forests are quite efficient during test time, we compare our method against affine registration to evaluate the accuracy of the initial set of

correspondences. Additionally we compare our method against a conventional organ label classification forest (LF) and a coordinate regression forest (RegF).

We use a dataset of 150 abdominal CT scans acquired from different subjects. The 3D scans have an in-plane resolution of 512×512 with a number of slices between 238 and 1061. Voxel sizes vary from 0.55 to 0.82 with a slice spacing ranging from 0.4 to 0.8 mm. Manual organ segmentations of the liver, spleen, kidneys, and pancreas are provided by clinical experts.

Given a test image that we would like to segment, we treat the remaining 149 images as atlases. The MALP setting would then be as follows:

- Select a subset of the most similar atlases as measured globally by SSD similarity between down-sampled versions of the atlases and the test image.
- The next step is obtaining a label prediction Lp_a from each atlas a. For **LF**, we simply apply the atlas forest on the test image to obtain Lp_a directly. For **affine registration**, all the images are affinely aligned to a template space. The labels from the atlas are then transferred to the test image based on the one-to-one voxel correspondences. For **RegF**, applying the atlas regression forest yields correspondences between the coordinates of the atlas and the test image. The labels are then transferred from the atlas to the test image. Lastly, applying an atlas **SVF** on the test image yields correspondences between the atlas and the test image on a supervoxel level. Each supervoxel from the atlas has an organ label which is obtained via majority voting from the organ labels of its voxels. The supervoxel-level organ labels are then transferred from the atlas to the supervoxels of the test image.
- The final labelling Lp of the test image is then obtained by fusing Lp_a from all the atlases via a voxel-wise majority vote.

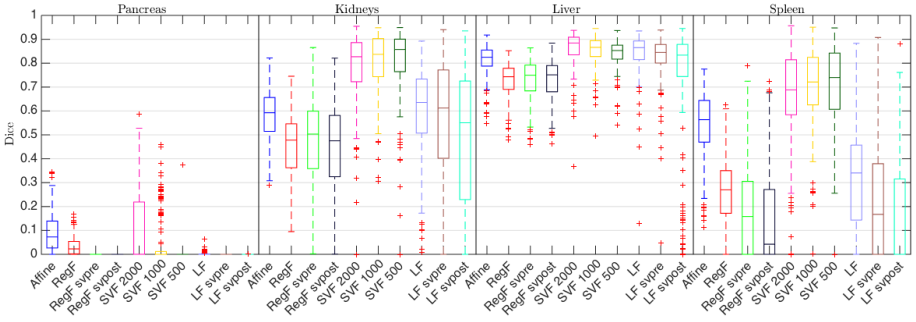

Fig. 3. Dice overlap using the 20 closests atlases to perform MALP. Results for affine registration (Affine), SVF, RegF, RegF svpre, RegF svpost, LF, LF svpre, and LF svpost. We see that SVF scores a higher dice overlap, especially for the kidneys and spleen. We also note that using supervoxels as a post-processing step with RegF and LF does not improve the prediction result.

To test whether using supervoxels can have an influence on the labelling obtained from RegFs and LFs, we apply a post-processing step – by assigning to each supervoxel the most frequent label of its voxels – either on the predictions Lp_a of each atlas before fusing (svpre) or directly on Lp after the fusion (svpost). We use 2000 supervoxels. In addition, for SVFs, we test with with 500, 1000, and 2000 supervoxels on average per image. Fig. 3 shows the Dice overlap from the different methods using the closest 20 atlases.

All the forests (LF, SVF, RegF) are trained with the same parameters: 5 trees –as not much difference has been observed from using 1 to 5 trees [14]–, maximum depth 32, minimum samples 4, $n_{\text{thresholds}} = 15$, $n_f = 500$, and with the images down-sampled to a $2 \times 2 \times 2mm^3$ spacing. With an implementation in C++, generating supervoxels on full resolution images takes around 30-50 seconds per image on a single machine with core i7 @ 3.40 GHz with 16 GB memory. For SVFs, training on a $160 \times 160 \times 93$ volume takes $\sim$ 4 mins/tree while testing takes $\sim$ 4s with a pre-processing time of $\sim$ 10s.

4 Discussion and Conclusion

In this paper we propose a method for estimating correspondences between images on a supervoxel level using classification forests. The advantage of our approach is that it does not rely on the availability of prior organ annotations. Training a random forest using automatically generated supervoxels as class labels allows training on unlabelled images. Qualitative evaluation of the estimated correspondences in a simple multi-atlas propagation setting demonstrate the potential of using SVFs for estimating correspondences. We do not apply any further post-processing to improve the segmentation, such as graph-cuts, which is what is typically done in some state-of-the-art methods for segmenting abdominal datasets [13, 11]. Random forests are extremely efficient during test time making them an attractive option to use for estimating correspondences in large datasets.

In addition, results seem to indicate that using an SVF to propagate labels from an atlas to a target image yields a higher prediction accuracy than using traditional random forests, such as a LF or a RegF. One possible reason might be that LFs have difficulty learning features to distinguish between one organ vs another, if an organ, for example the liver, covers a wider span of contextual appearance features due its size. Whereas RegFs ignore organ boundaries and will mix voxels from organs with those of the background. On the other hand, SVFs offer a nice balance between locality and tissue type consistency via the use of supervoxels.

The current supervoxel segmentation is not optimal when using a small number of supervoxels that do not adhere perfectly to the boundaries of the underlying ground truth segmentation. This is especially true for the pancreas. Computing Dice overlaps between the ground truth organ labels and the their supervoxelised version –obtained by assigning to each supervoxel the majority vote of the ground truth label of its voxels– yields for 150 images: pancreas 0.641 ± 0.138,

kidneys 0.927 ± 0.061, liver 0.935 ± 0.022, and spleen 0.908 ± 0.054. Future work would include investigating more appropriate supervoxel segmentation and hierarchical representations. Moreover, it would be interesting to exploit mutual correspondences as it is possible to obtain them by training independently on both images, then testing on each other and keeping only the mutual correspondences. For purposes of evaluation such an approach would require a more sophisticated label propagation scheme which we do not adopt here.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC superpixels. No. EPFL-REPORT-149300 (June), 15 (2010)
2. Breiman, L.: Random forests. Machine learning pp. 5–32 (2001)
3. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and regression trees. CRC press (1984)
4. Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54(2), 940–954 (2011)
5. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. Learning 7, 81–227 (2011)
6. 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)
7. 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)
8. Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain mri segmentation combining label propagation and decision fusion. NeuroImage 33(1), 115–126 (2006)
9. Lucchi, A., Smith, K., Achanta, R., Knott, G., Fua, P.: Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. IEEE transactions on medical imaging 31(2), 474–86 (Feb 2012)
10. 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)
11. 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)
12. Wang, H., Yushkevich, P.A.: Multi-atlas segmentation without registration: a supervoxel-based approach pp. 535–542 (2013)
13. Wolz, R., Chu, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D.: Automated abdominal multi-organ segmentation with subject-specific atlas generation. Medical Imaging, IEEE Transactions on 32(9), 1723–1730 (2013)
14. Zikic, D., Glocker, B., Criminisi, A.: Encoding atlases by randomized classification forests for efficient multi-atlas label propagation. Medical image analysis (Jul 2014)
15. Zitova, B., Flusser, J.: Image registration methods: a survey. Image and vision computing 21(11), 977–1000 (2003)