Selective Labelling: identifying representative sub-volumes for interactive segmentation

Imanol Luengo*, Mark Basham, and Andrew P. French

School of Computer Science, University of Nottingham, Nottingham, UK, NG8 1BB
{imanol.luengo,andrew.p.french}@nottincham.ac.uk
Diamond Light Source Ltd, Harwell Science & Innovation Campus,
Didcot, UK, OX11 0DE
mark.basham@diamond.ac.uk

Abstract. Automatic segmentation of challenging biomedical volumes with multiple objects is still an open research field. Automatic approaches usually require a large amount of training data to be able to model the complex and often noisy appearance and structure of biological organelles and their boundaries. However, due to the variety of different biological specimens and the large volume sizes of the datasets, training data is costly to produce, error prone and sparsely available. Here, we propose a novel Selective Labelling algorithm to overcome these challenges; an unsupervised sub-volume proposal method that identifies the most representative regions of a volume. This massively-reduced subset of regions are then manually labelled and combined with an active learning procedure to fully segment the volume. Results on a publicly available EM dataset demonstrate the quality of our approach by achieving equivalent segmentation accuracy with only 5% of the training data.

Keywords: unsupervised, sub-volume proposals, interactive segmentation, active learning, affinity clustering, supervoxels

1 Introduction

Automatic segmentation approaches have yet to have an impact in biological volumes due to the very challenging nature, and wide variety, of datasets. These approaches typically require large amounts of training data to be able to model the complex and noisy appearance of biological organelles. Unfortunately, the tedious process of manually labelling large volumes with multiple objects, which takes days to weeks for a human expert, makes it infeasible to generate reusable and generalizable training data. To deal with this absence of training data, several semi-automatic (also called *interactive*) segmentation techniques have been proposed in the medical imaging literature. This trend has been rapidly growing over the last few years due to the advances in fast and efficient segmentation techniques. These approaches have been used to interactively segment a wide variety of medical volumes, such as arbitrary medical volumes[1] and organs[2].

* Corresponding Author

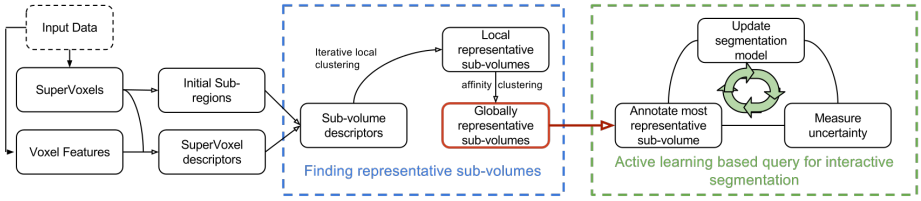

Fig. 1. Overview of our proposed pipeline

However, segmenting large biological volumes with tens to hundreds of organelles requires much more user interaction for which current interactive systems are not prepared. With current systems, an expert would need to manually annotate parts of most (or even all) the organelles in order to achieve the desired segmentation accuracy. To deal with the absence of training data and assist the human expert with the interactive segmentation task we propose a Selective Labelling approach. This consists of a novel unsupervised sub-volume[1] proposal method to identify a massively reduced subset of windows which best represent all the textural patterns of the volume. These sub-volumes are then combined with an active learning procedure to iteratively select the next most informative sub-volume to segment. This subset of small regions combined with a smart region-based active learning query strategy preserve enough discriminative information to achieve state-of-the-art segmentation accuracy while reducing the amount of training data needed by several orders of magnitude.

The work presented here is inspired by the recent work of Uijlings et al.[3] (Selective Search) which extracts a reduced subset of multi-scale windows for object segmentation and has been proved to increase the performance of deep neural networks in object recognition. We adapt the idea of finding representative windows across the image under the hypothesis that a subset of representative windows have enough information to segment the whole volume. Our approach differs from Selective Search in the definition of what *representative windows* are. Selective Search tries to find windows that enclose objects, and thus, they apply a hierarchical merging process over the superpixel graph with the aim of obtaining windows that enclose objects. Here, we adopt a completely different definition of *representative windows* by searching for a subset of fixed-sized windows along the volume that best represent the textural information of the volume. This provides a reduced subset of volume patches that are easier to segment and more generalizable. Active learning techniques have been applied before in medical imaging[4][5], but they have been focused on querying the most uncertain voxels or slice according to the current performance of the classification model in single organ medical images. Our approach differs from other active learning approaches in medical imaging by: (1) It operates in the supervoxel space, making the whole algorithm several orders of magnitudes faster. (2) It first extracts a subset of representative windows which are used to loop the

[1] The terms *sub-volume* and *window* will be used interchangeably along the document.

active learning procedure. Training and querying strategy is only applied in a massively reduced subset of data, reducing computational complexity. (3) The queries for the user are fixed-sized sub-volumes which are very easy to segment with standard graphcut techniques. To summarize, the main contributions of the current work can be listed as follows:

1. A novel representative patch retrieval system to select the most informative sub-volumes of the dataset.
2. A novel active learning procedure to query the window that would maximize the model's performance.
3. Our segmentation framework, used as an upper bound measure, achieves similar performance to [6] while being much faster.

2 Segmentation framework

To be able to segment large volumes efficiently, we adopt the supervoxel strategy introduced by Lucchi et al. [6] to segment mitochondria from Electron Microscopy (EM) volumes. Supervoxels consist of a group of neighbouring voxels in a given volume that share some properties, such as texture or color. Each of the voxels of the volume belong to exactly one supervoxel, and by adopting the supervoxel representation of a dataset, the complexity of a problem can be reduced two or three orders of magnitude. A supervoxel graph is created by connecting each supervoxel to its neighbours (the one it shares boundary with). Then, we extract local textural features from each voxel of the volume:

$$\mathbf{f} = \{G^{\sigma_1}, G_x^{\sigma_1}, G_y^{\sigma_1}, G_z^{\sigma_1}, G^{\sigma_5}, G_x^{\sigma_5}, G_y^{\sigma_5}, G_z^{\sigma_5}\}, \tag{1}$$

where G^{σ_x} represents the input volume convolved by a Gaussian filter of $\sigma = x$ and subscript indicates directional derivatives. Supervoxel descriptors $\boldsymbol{\phi}_k$ (for supervoxel k) are then computed as Sigma Set[7] features of all the voxels belonging to them. These descriptors map the supervoxel covariance to a Euclidean space which has been proved to be a very efficient and robust for classification in sub-cellular volumes in [8]. To improve the accuracy and the robustness of the supervoxel descriptors, contextual information is added by appending for each supervoxel the mean $\boldsymbol{\phi}$ of all its neighbours:

$$\boldsymbol{\psi}_k = \left[\boldsymbol{\phi}_k\,,\ \frac{1}{m}\sum_{i\in\mathcal{N}(k)}\boldsymbol{\phi}_k\right] \tag{2}$$

Segmentation is then formulated as a Markov Random Field optimization problem defined over the supervoxel graph with labels $\mathbf{c} = \{c_i\}$:

$$E(\mathbf{c}) = \sum_{s_i\in SV} E_{data}(s_i, c_i) + \beta \sum_{(s_i,s_j)\in\mathcal{N}} E_{smooth}(c_i, c_j). \tag{3}$$

Here, the data fidelity term E_{data} is defined as the negative log likelihood of the output of an Extremely Random Forest[9] (ERF), with $T = 100$ trees, trained

on the supervoxel features $\boldsymbol{\psi}_k$. The pairwise potential E_{smooth} is also learnt from data (similar to [6]) with another ERF by concatenating the descriptors of every pair of adjacent supervoxels with the aim of modelling the *boundariness* of a pair of supervoxels. We refer the reader to [6] for more information about this segmentation model as it is used only as an upper bound and is out of the scope of this paper improving the framework.

3 Finding representative sub-volumes

Biological volumes are usually very large (here for example $1024 \times 768 \times 330$). In order to efficiently segment them, we provide a framework to extract most representative sub-volumes which can then be used to segment the rest of the volume. We start by defining a fixed size V_s for the sub-volumes, set empirically to preserve information whilst being easy to segment. In this work, we set $V_s = [100, 100, 10]$. Considering every possible overlapping window centered at

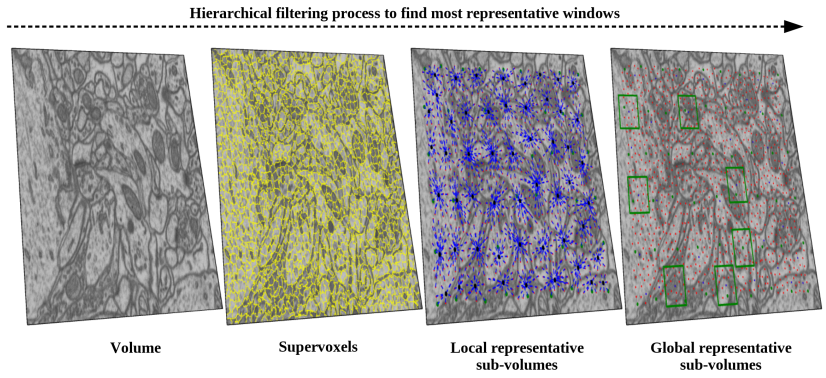

Fig. 2. Overview of the window proposal method. For visualization purposes a 2D slice is shown, but every step is performed in 3D.

each voxel of the volume would generate too many samples (around $200M$ voxels). Thus, we start by considering the set of proposed windows $w \in \mathcal{W}$ from N windows centered at each of the supervoxels of the image, as we already know these regions are likely to have consistent properties. We extract $10 \times 10 \times 10$ supervoxels which reduces the amount of windows by 3 orders of magnitude to roughly 200K. Next, in order to extract representative regions from the image we first need to define how to describe a region. To do so, we first cluster all the supervoxel descriptors $\boldsymbol{\phi}_k$ in $B = 50$ bins to assign a *texton* to each supervoxel. The regional descriptor $\mathbf{r}_k$, assigned to the window proposal w_k centered at supervoxel k, is the $\ell 1$-normalized histogram of supervoxel textons in that window. Thus, $\mathbf{r}_k$ encodes the different textural patches and the proportion of each of them present in each window. The descriptor is rotationally invariant and very powerful discriminative descriptor for a region.

3.1 Grouping similar nearby sub-volumes

Once sub-volume descriptors are extracted, we perform a second local clustering. Similar to SLIC to create supervoxels but to cluster together nearby similar sub-volumes. To do so, we first sample a grid of V_s cluster centers $C_i \in \mathcal{C}$ uniformly across the volume and assign them to their nearest window w_k. For each window we use their position $\mathbf{p}_k$ in the volume and their descriptor $\mathbf{r}_k$. Then, the local k-means clustering iterates as follows:

1. **Assign each sub-volume to their nearest cluster center.** For each cluster C_i compute the distance to each of the windows in a neighbourhood. The neighbourhood is set to $2 \times V_s = [200, 200, 20]$.

$$d(C_i, k) = \left\| \mathbf{p}_{C_i} - \mathbf{p}_k \right\|_2 + \frac{\lambda}{\sqrt{2}} \left\| \sqrt{\mathbf{r}_{C_i}} - \sqrt{\mathbf{r}_k} \right\|_2 \tag{4}$$

   where the first term represents the standard Euclidean spatial distance between windows and the second term is the Hellinger distance that measures the difference in appearance of the windows. Each window w_k is assigned to the neighbouring cluster C_i(label L_k) that minimizes the above distance.
2. **Update cluster centers.** The new cluster center is the assigned the window to minimizes the sum of differences with all the other windows, or in other words, the window that best represents all the others assigned to the same cluster:

$$C_i = argmin_{k \in \{k | L_k = i\}} \sum_{j \in \{j | L_j = i\}} d(k, j) \tag{5}$$

The above update is very efficient and clusters nearby and similar windows into a even smaller set. After 5 iterations of the above procedure, the number of proposal windows $w_k \in \mathcal{W}$ is reduced from $200K$ to 3500 by only considering the windows that best describe their neighbouring windows $w_{C_i} \in \mathcal{W}$. Let us refer to this reduced set of windows as $\mathcal{R}$.

3.2 Further refining window proposals

After filtering the window proposals that best represent their local neighbourhood, still a large number of possible sub-volumes remain. To further filter the most representative regions from $w_k \in \mathcal{R}$ we apply a affinity propagation based clustering[10]. Affinity propagation clustering is a message-passing clustering that automatically detects *exemplars*. The inputs for affinity clustering consist of an *affinity matrix* as the connection weights between data points and the *preference* of assigning each of the data points as *exemplars*. Then, through an iterative message-passing procedure the affinity propagation refines the weights between data points and the *preferences* until the optimal (and minimal) subset of *exemplars* is found. After local representative regions are extracted from section 3.1, the pairwise similarity between all the remaining regions $w_k \in \mathcal{R}$ is extracted as

$$a(i, j) = intersection(\mathbf{r}_i, \mathbf{r}_j) \tag{6}$$

to form the $M \times M$ affinity matrix A, where $A_{i,j} = a(i,j)$ is the similarity (only in appearance, measured by the *intersection* kernel) between all pairs of windows w_i and w_j. The *preference* vector P is set to a constant weighted by the ℓ_∞ norm of the appearance vector $P_i = \gamma\,(1 - \|\mathbf{r}_i\|_\infty)$. The ℓ_∞ norm of a vector returns the maximum absolute value of the vector. For a ℓ_1 normalized histogram is a good measure of how spread the histogram is. Thus, the weight $(1-\ell_\infty)$ will encourage windows that contain wider variety of textural features to be selected. This is a desired feature, since we aim to extract a very small subset of window proposals for the whole volume, we would expect them to represent all the possible textural features of the volume or if not the training stage will fail to model unrepresented features. After the affinity propagation clustering, we now have a manageable set of < 100 sub-volumes which together represent the global appearance of the whole volume. Let us denote this final subset of proposals as $\mathcal{P}$.

4 Querying the next most informative sub-volume

The active learning cycle starts once a minimal representative set of sub-regions $\mathcal{P}$ has been extracted and at least 1 window (containing both foreground and background) has been segmented. From there, the ERF model from section 2 is trained and used to predict the labels of all the supervoxels belonging to all the windows in $\mathcal{P}$. Here, we average the probabilistic prediction of all the trees $t \in T$ of the ERF in order to model the probability of a supervoxel to belong to foreground or background. The uncertainty of its prediction is then estimated as the entropy. Then, the average uncertainty of all the supervoxels U_s in a window $w_k \in \mathcal{P}$ is defined as the average uncertainty in the predictions of all the supervoxels contained in that window. Similarly, the average uncertainty of *boundariness* U_e of all connected pair of supervoxels in a window is extracted from the other ERF trained to identify this property. The average window uncertainty is then defined as $U_w = U_s + \beta\, U_e$. The window with larger average uncertainty is selected as the next sub-volume to be segmented. As all the windows have been previously reduced to a minimal subset, the query strategy is very efficient and is able to return a globally representative sub-volume that would maximize the performance of the ERF classifier.

5 Experiments

In our experiments we used the publicy available EM dataset[2] used in [6]. The data set consists of a $5\times5\times5\mu m$ section taken from the CA1 hippocampus region of the brain. Two $1024 \times 768 \times 165$ volumes are available where mitochondria are manually annotated (one for training and the other one for testing). We first validate the results of our segmentation pipeline by using one of the volumes for training while the other for testing. Table 1 shows results of different stages of

[2] `http://cvlab.epfl.ch/data/em`

our segmentation pipeline: (1) $\mathbf{ERF}_{raw}$ evaluates only the prediction of the ERF trained in the supervoxel features, (2) $\mathbf{ERF}_{nh}$ is the prediction of the ERF after aggregating neighbouring supervoxel features, (3) $\mathbf{MRF}_{nh}$ is (2) refined with a contrast-sensitive MRF and (4) $\mathbf{MRF}_{learnt}$ is the full model with learned unary and pairwise potentials. Our model, used as an upper bound of the *maximum*

Table 1. Performance of our segmentation pipeline in the testing dataset

	$\mathbf{ERF}_{raw}$	$\mathbf{ERF}_{nh}$	$\mathbf{MRF}_{nh}$	$\mathbf{MRF}_{learnt}$
Accuracy	0.975	0.984	0.987	0.991
DICE coefficient	0.751	0.825	0.851	0.871
Jaccard index	0.601	0.702	0.743	0.780

achievable accuracy of the following experiment. It has similar segmentation performance to the one reported in [6], while being much faster (15 minutes of processing and training time vs 9 hours).

Table 2 shows a benchmark of the quality and descriptive power of a reduced subset of our extracted windows. To evaluate the quality of our extracted windows, we simulate different user patterns. *Random User* will define the behaviour of a user selecting n random patches for training across the training volume. *Random Oracle* will select n random patches for training centered in a supervoxel that belongs to mitochondria (thus, assumes ground truth is known and simulates the user clicking in different mitochondria). *Selective Random* simulates a user choosing n windows at random from a reduced subset of windows $w_k \in \mathcal{P}$ obtained using our algorithm. And *Selective Labelling* will select the first window at random from $w_k \in \mathcal{P}$ (containing both background and foreground) while the next $n-1$ will be selected by our active learning based query strategy. All different patterns are trained only on the selected windows of the training volume (with the full model) and tested in the whole testing volume. The 3 random patterns are averaged from 100 runs. It can be seen that our extracted

Table 2. DICE coefficient of the simulated retrieval methods. Percentages indicate fractions of total training data.

	Random User	*Random Oracle*	*Selective Random*	*Selective Labelling*
3 sub-volumes ($< 1\%$)	0.305	0.671	0.652	0.788
5 sub-volumes (1%)	0.533	0.736	0.740	0.792
10 sub-volumes (2%)	0.608	0.762	0.761	0.810
30 sub-volumes (5%)	0.691	0.805	0.803	0.841

windows without the active learning achieve similar performance to the random

oracle (which assumes ground truth is known). This proves the quality of our windows as our unsupervised method is able to represent properly all the textural elements of the volume. With the active learning, our method outperforms all the others and is able to obtain similar performance to the baseline trained in the whole volume (table 1) with much fewer training data (up to 5%).

6 Conclusions and future work

We have presented a fully unsupervised approach to select the most representative windows of the volume, which combined with a novel active learning procedure obtain similar accuracy than fully automatic methods by using only 5% of the data for training. The presented segmentation pipeline achieves similar performance to the state-of-the-art in a publicly available EM dataset, while being much faster and efficient. The results demonstrate that with the assistance of the proposed algorithm, a human expert could segment large volumes much faster and easier. It also makes the segmentation task much more intuitive by giving the user small portions of the volume, which are much easier to annotate. Extension to multi-label interactive segmentation is straight forward as all the methods here presented are inherently multi-label.

References

1. Peter Karasev, Ivan Kolesov, Karl Fritscher, Patricio Vela, Paul Mitchell, and Allen Tannenbaum. Interactive medical image segmentation using pde control of active contours. *Medical Imaging, IEEE Transactions on*, 2013.
2. Reinhard Beichel et al. Liver segmentation in ct data: A segmentation refinement approach. *3D Segmentation in The Clinic: A Grand Challenge*, 2007.
3. Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 2013.
4. Andrew Top, Ghassan Hamarneh, and Rafeef Abugharbieh. Active learning for interactive 3d image segmentation. In *MICCAI 2011*. 2011.
5. Andrew Top et al. Spotlight: Automated confidence-based user guidance for increasing efficiency in interactive 3d image segmentation. In *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*. 2010.
6. Aurélien Lucchi et al. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *Medical Imaging, IEEE Transactions on*, 31(2):474–486, 2012.
7. Xiaopeng Hong, Hong Chang, Shiguang Shan, Xilin Chen, and Wen Gao. Sigma set: A small second order statistical region descriptor. In *CVPR 2009*, 2009.
8. Imanol Luengo, Mark Basham, and Andrew P French. Fast global interactive volume segmentation with regional supervoxel descriptors. In *SPIE Medical Imaging*, pages 97842D–97842D. International Society for Optics and Photonics, 2016.
9. Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.
10. Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.