Boosting Hand-Crafted Features for Curvilinear Structure Segmentation by Learning Context Filters*

Roberto Annunziata[1,**], Ahmad Kheirkhah[2], Pedram Hamrah[2], and Emanuele Trucco[1]

[1] School of Computing, University of Dundee, Dundee, UK
r.annunziata@dundee.ac.uk
[2] Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, USA

Abstract. Combining hand-crafted features and learned filters (i.e. feature boosting) for curvilinear structure segmentation has been proposed recently to capture key structure configurations while limiting the number of learned filters. Here, we present a novel combination method pairing hand-crafted appearance features with learned context filters. Unlike recent solutions based only on appearance filters, our method introduces *context information* in the filter learning process. Moreover, it reduces the potential redundancy of learned appearance filters that may be reconstructed using a combination of hand-crafted filters. Finally, the use of k-means for filter learning makes it fast and easily adaptable to other datasets, even when large dictionary sizes (e.g. 200 filters) are needed to improve performance. Comprehensive experimental results using 3 challenging datasets show that our combination method outperforms recent state-of-the-art HCFs and a recent combination approach for both performance and computational time.

1 Introduction

Clinical research has shown that morphometric properties of curvilinear structures in the human body, e.g. blood vessels, are associated with a wide range of sub-clinical (e.g. atherosclerosis) and clinical cardiovascular diseases (e.g. diabetes, hypertension, stroke) [1]. Automated systems would allow cost-effective large screening programs aimed at early diagnosis.

To measure automatically morphometric properties such as tortuosity [2], these structures have to be segmented accurately. Many solutions have been proposed to cope with the multiple challenges involved, including low signal-to-noise ratio at small scales, confounding non-target structures, non-uniform illumination and complex configurations [3–11]. Most approaches are based on a local

* This research was supported by the EU Marie Curie ITN 316990 "REVAMMAD". The authors are grateful to Amos Sironi (CVlab, EPFL) for providing the VC6 and BF2D datasets and to Max Law (Western University) for the OOF code.

** Corresponding author.

N. Navab et al. (Eds.): MICCAI 2015, Part III, LNCS 9351, pp. 596–603, 2015.
DOI: 10.1007/978-3-319-24574-4_71

tubularity measure estimated via hand-crafted features (henceforth, HCFs) [3, 4, 6, 7, 10], or learned from training data [5, 9, 11]. Although HCFs can be fast, they may rely on assumptions violated in some cases (e.g. at bifurcations and crossing points). For this reason, Honnorat et al. [7] use a graphical model after HCF-based tubularity estimation to improve guide-wire tracking during cardiac angioplasty interventions. Other methods are based on fully learned architectures capturing key configurations on training data, but can be demanding computationally since they typically require more filters than efficient HCFs [3, 6]. Recently, Rigamonti et al. [8] proposed a novel, hybrid approach combining HCFs with learned appearance filters (i.e. feature boosting) to exploit the efficiency of fast HCFs while limiting the amount of learned filters. It employs sparse coding (henceforth, SC) to learn 9 appearance filters and combines them with state-of-the-art HCFs. Although this method outperforms state-of-the-art approaches such as [3, 6], our work is motivated by limitations including the amount of false positives far from target structures and modest success in segmenting fragmented or weakly connected structures such as corneal nerve fibres and neurons (Section 3). Integrating *context* information with *appearance* features has been recently found to address these shortcomings, albeit at an extra computational cost as multiple discriminative models are learned sequentially [12].

In this paper, we propose a novel combination approach in which efficient HCFs capturing the *appearance* are paired with learned *context* filters. Overall, our method has three key advantages over recently proposed methods (e.g. [8]): 1) it includes context information recently found to improve results over appearance-only frameworks [12] while still learning a *single* discriminative model; 2) it implicitly reduces or eliminate the redundancy of learned filters; 3) it is fast and easily adaptable to other datasets, even when large dictionary sizes (e.g. 200 filters) are needed to improve performance.

2 Proposed Approach

Unsupervised Filter Learning. k-means clustering has been successfully used for filter/dictionary learning, often competing with more complex state-of-the-art methods [13]. Although k-means is not designed to learn "distributed" representations, like sparse coding or independent component analysis, experiments suggest that k-means tends to discover sparse projections of the data [13] with sufficiently large numbers of training patches given the patch size and patch-level whitening to remove correlations between nearby pixels. Because of this and given the large amount of our training data (Section 3), we employ k-means clustering to learn filters instead of algorithms explicitly designed to obtain sparse representations like SC used in [8]. We use the k-means filter learning algorithm described in [13] which we summerize succinctly.

Our goal is to learn a dictionary $\mathbf{D} \in \mathbb{R}^{n \times k}$ of k vectors so that a data vector $\mathbf{x}^{(i)} \in \mathbb{R}^n, i = 1, \ldots, m$ can be mapped to a code vector that minimizes the reconstruction error. Before running the learning algorithm we normalize the brightness and contrast of each input data point (i.e. patch) $\mathbf{x}^{(i)}$. Then, we apply patch-level

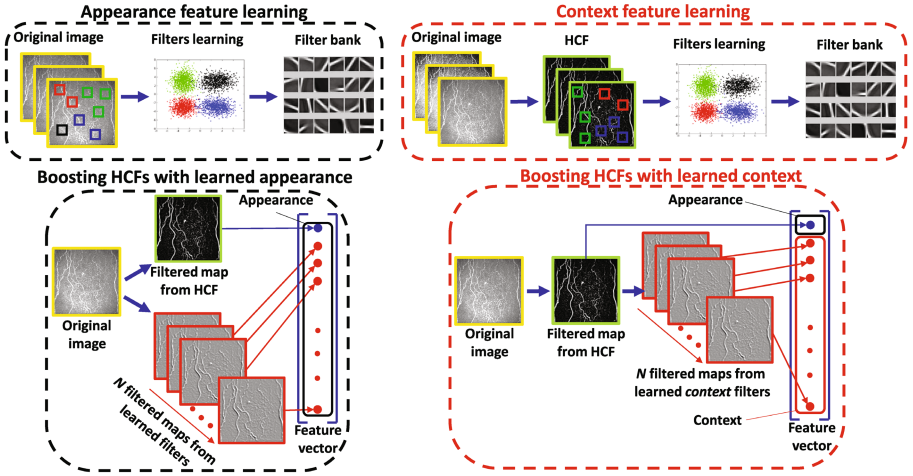

Fig. 1. Boosting HCFs with learned appearance and context filters. In the former, learned filters are applied to original image; HCF maps are used in the latter.

whitening through the ZCA transform so that $\mathbf{x}_{ZCA}^{(i)} = \mathbf{V}(\mathbf{\Sigma}+\epsilon_{ZCA}\mathbf{I})^{-1/2}\mathbf{V}^T\mathbf{x}^{(i)}$, where $\mathbf{V}$ and $\mathbf{\Sigma}$ are computed from the eigenvalue decomposition of the data points covariance $\mathbf{V\Sigma V}^T = \text{cov}(\mathbf{X})$, and ϵ_{ZCA} is a small constant (Section 3) controlling the trade-off between whitening and noise amplification. After pre-processing the patches, we solve the optimization problem

$$\underset{\mathbf{D},\mathbf{c}}{\text{argmin}} \sum_i \left\| \mathbf{D}\mathbf{c}^{(i)} - \mathbf{x}_{ZCA}^{(i)} \right\|_2^2 \tag{1}$$

subject to $||\mathbf{c}^{(i)}||_0 \leq 1, \forall i = 1, \ldots, m$ and $||\mathbf{d}^{(j)}||_2 = 1, \forall j = 1, \ldots, k$, where $\mathbf{c}^{(i)}$ is the code vector related to input $\mathbf{x}_{ZCA}^{(i)}$, and $\mathbf{d}^{(j)}$ is the j-th column of the dictionary $\mathbf{D}$.

Combining appearance filters learned through k-means (or SC) with HCFs leads to the combination approach proposed in [8] based on appearance-only features (see Figure 1 left).

Learning Context Filters. Hand-crafted and learned appearance features are designed to capture object-specific properties. In the case of curvilinear structures, they detect characteristics that make them *appear* as ridge-, vessel- or tube-like shapes. However, appearance features do not take into account specific inter-object relationships, *context* information that has been found recently to improve performance significantly in segmentation tasks over methods employing appearance features only. One well-known method including context information is auto-context [12], which learns multiple discriminative models sequentially. This imposes an extra computational cost over learning a single discriminative model, as in traditional methods based on features representing object pixels and a single classifier to infer their labels. While an extra computational cost may have little impact on training/testing time for applications

involving small datasets or small images, it may become impractical with large volumes of images data. This is the case, for instance, with large screening programs based on images, e.g., diabetic retinopathy [1]. For this reason, our goal is to include context information in a learning method without increasing the computational cost with respect to the solution in [8]. This is achieved by learning a *single* discriminative model which takes as input both *appearance* (i.e. likelihood computed on the original image) and, unlike the method in [8], *context information* (i.e. relations between objects). To model appearance, we employ the fast Optimally Oriented Flux (OOF) feature [6], shown to outperform other HCFs on the datasets we use for our tests [8]. We include context information by learning context filters to be used in combination with the HCFs in a new hybrid model to segment curvilinear structures. Learning context filters has two clear advantages: 1) including high-level information in a hybrid framework; 2) high efficiency and adaptability since convolution with filter banks is very fast on standard computers. In addition, our proposed method has a key advantage over methods learning appearance filters (e.g. [8]): it implicitly eliminates, or reduces significantly, the redundancy of learned filters. In fact, learned appearance filters may be reconstructed through a combination (linear or non-linear) of the HCFs already used to model appearance, thus reducing the discrimination power of the feature set. Figure 1 shows the difference between the proposed approach (right) and the combination method proposed in [8] (left). Notice, while appearance filters are learned on the same layer where HCFs are applied (original image), context filters are learned on a *different* layer (i.e. after HCFs are applied).

Description Vector and Supervised Classification. Learned filters are applied to the input image to compute multiple feature maps efficiently using correlation:

$$\mathbf{L}^{(j)} = \mathbf{D}^{(j)} \circ \mathbf{I}_n, \tag{2}$$

where $\mathbf{D}^{(j)}$ is the j-th learned filter, $\mathbf{I}_n$ is the normalized input image (i.e., zero mean and unit standard deviation) and the $\circ$ symbol denotes correlation. We have also experimented with normalizations of the input image at patch level (including whitening), measuring the squared distance between the pre-processed patch and each filter; experiments (omitted given space limits) show that these normalizations, although important during the filter learning procedure, do not improve performance significantly and increase the computational cost. Thus, for each image location (u, v), we construct the following description vector:

$$[\underbrace{\mathbf{OOF}(u,v)}_{\textbf{appearance}}, \underbrace{\mathbf{L}^{(1)}(u,v), \ldots, \mathbf{L}^{(N)}(u,v)}_{\textbf{context}}]^T, \tag{3}$$

including appearance and learned context features ($N \leq 200$ in our experiments). We then apply a Random Decision Forest to classify each pixel. Centreline detection is obtained using Canny-like non-maxima suppression on the tubularity map. Local orientation is estimated using OOF.

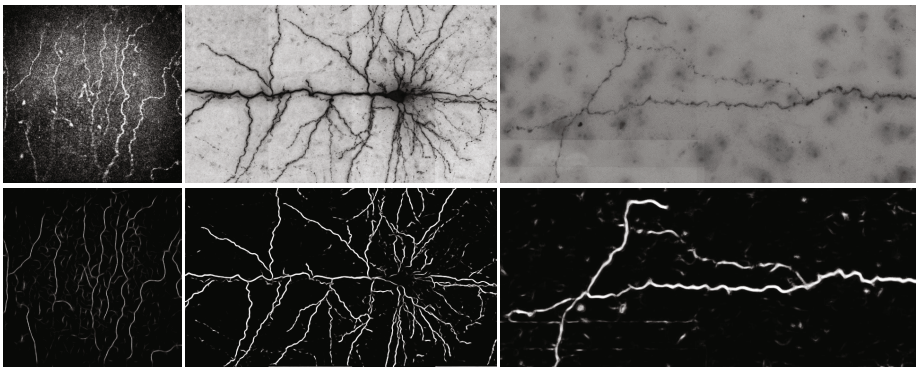

Fig. 2. Original images (top) and tubularity maps (bottom) obtained with our approach on IVCM (left), BF2D (centre), VC6 (right) datasets.

3 Experiments

Datasets. We validate our combination method using 3 datasets including low and high resolution images of corneal nerve fibres and neurons, showing very diverse images of curvilinear structures as shown in Figure 2.

IVCM [2] is a dataset of 100 384×384 confocal microscopy images capturing corneal nerve fibres with different grades of tortuosity, annotated by the clinical authors. Low resolution, non-uniform and poor contrast, tortuosity, and fibre fragmentation make this dataset particularly challenging. Following [11], we introduce a tolerance factor ρ for centreline detection: a predicted centreline point is considered a true positive if it is at most ρ pixels distant from a ground truth centreline point ($\rho = 2$ pixels [11]). Random sub-sampling is used to test performance on this dataset, using 50 images for training and the rest for testing in each cross-validation runs.

The **BF2D** dataset [8] consists of two minimum intensity projections of bright-field micrographs that capture neurons, annotated by an expert. The images have a high resolution (1024×1792 and 768×1792) but a low signal-to-noise ratio; the dendrites often appear as point-like structures easily mistaken for noise. We adopted the same dataset partition as in [8].

The **VC6** dataset [8] shows dendritic and axonal subtrees from one neuron in the primary visual cortex. It consists of three images obtained computing minimum intensity projections of 3D images, with numerous artifacts and poor contrast, hence challenging for automatic segmentation. We adopted the same dataset partition as in [8], selecting two images for training, and the third for testing. For all the datasets we averaged results over 10 random trials.

Experimental Setup. First, images are normalised to have zero mean and unit standard deviation. When HCFs are used as baselines, all parameters are tuned separately for each dataset to achieve best performance. To reduce the number of parameters to be optimised over datasets and test generalization, we fixed the OOF parameters range, the whitening parameter ϵ_{ZCA} and filters size in our filter banks to the same for *all* datasets. For OOF, we set $\sigma =$

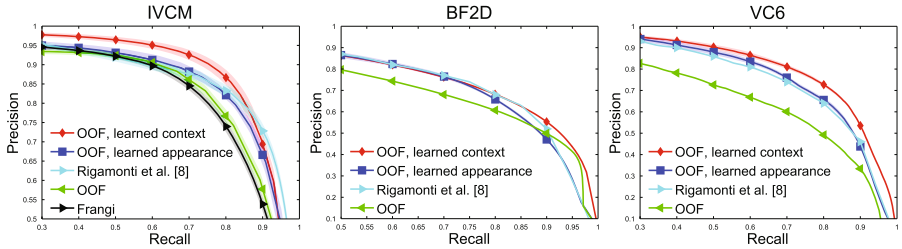

Fig. 3. Precision-recall curves for pixel-level classification. Shaded color bands represent 1 standard deviation of the results from individual runs.

$\{2, 3, 4\}$ (Eq. (8) [6]) and $R = \{2, 3, 4\}$ (Eq. (5) [14]). ϵ_{ZCA} was set to 0.001, considering the trade-off between noise amplification and filters sharpness [13]. Filters size was set to 11×11 pixels. Notice that the chosen patch size allows us to collect a sufficient number of patches to learn dictionaries, in agreement with the guidelines in [13][1]. We adopt filter banks of 100 filters (i.e., $N = 100$) for our method as good compromise between accuracy and speed. We used Random Decision Forests with 100 random trees to make predictions fast. We trained classifiers using the same number of positive and negative samples and priors estimated empirically. Experiments were run on a 8-core 64-bit architecture using MATLAB implementations.

We compare our method, combining appearance and context filters, with the one recently introduced in [8] based only on appearance. Therefore, we report results obtained with the original implementation of [8] using 9 appearance filters learned through SC ("*Rigamonti et al. [8]*" in Figure 3)[2]. We also report experiments using k-means to learn appearance filters ("*OOF, learned appearance*") and the same dictionary size we adopted to learn context filters.

Results and Discussion. As Figure 3 shows, our feature boosting method ("*OOF, learned context*") outperforms the baselines on all datasets, especially on IVCM and VC6. The variability of the datasets allows us to compare performance with very diverse data characteristics: with tortuous, fragmented structures and low signal-to-noise ratio (IVCM), our method shows a better precision for low recall values as it better segments fragmented structures; when structures have better contrast but point-like appearance (BF2D), our method shows higher precision at high recall values as it reduces the false positives due to point-like non-target structures and increases the connectivity; when dealing with complex non-target structures (e.g. blobs in VC6), our approach shows better performance from medium to high recall values, since it reduces the amount of false positives due to such structures.

Since our main goal is to compare learned context with learned appearance regardless of the learning algorithm used (k-means or SC), we evaluate the effect

[1] We selected 340,000 patches in the worst case represented by a single training image of the BF2D dataset. Notice that, in [13] 100,000 16×16 patches are considered sufficient.

[2] Original code at: `https://bitbucket.org/roberto_rigamonti/med_img_pc`.

Table 1. Effect of patch and dictionary size on the area under precision-recall curves (mean/standard deviation). Our method outperforms the baseline in all conditions.

IVCM	**Patch size (pixels)**			**Number of learned filters (N)**		
	11×11	15×15	21×21	10	100	200
OOF, learned context	0.8928/0.0056	0.9078/0.0043	0.9133/0.0052	0.8866/0.0033	0.8928/0.0056	0.8976/0.0037
OOF, learned appearance	0.8665/0.0114	0.8878/0.0059	0.8870/0.0102	0.8557/0.0069	0.8665/0.0114	0.8878/0.0039
	100 learned filters			**Patch size:** 11×11 **pixels**		

BF2D	**Patch size (pixels)**			**Number of learned filters (N)**		
	11×11	15×15	21×21	10	100	200
OOF, learned context	0.8057/0.0017	0.8010/0.0029	0.7948/0.0021	0.7966/0.0043	0.8057/0.0017	0.8054/0.0022
OOF, learned appearance	0.7881/0.0060	0.7888/0.0035	0.7908/0.0036	0.7857/0.0026	0.7881/0.0060	0.7824/0.0044
	100 learned filters			**Patch size:** 11×11 **pixels**		

VC6	**Patch size (pixels)**			**Number of learned filters (N)**		
	11×11	15×15	21×21	10	100	200
OOF, learned context	0.8272/0.0052	0.8295/0.0035	0.8069/0.0063	0.7911/0.0070	0.8272/0.0052	0.8368/0.0041
OOF, learned appearance	0.7905/0.0063	0.7904/0.0045	0.7809/0.0032	0.7460/0.0053	0.7905/0.0063	0.7905/0.0062
	100 learned filters			**Patch size:** 11×11 **pixels**		

of patch and dictionary size on the performance measured using Area Under Precision-Recall Curve (AUPRC) for both combination methods using k-means as learning method. As Table 1 shows, combining OOF with learned context (proposed here) outperforms the combination with learned appearance [8] in terms of AUPRC regardless of the chosen patch and dictionary size. Moreover, learning as little as 10 context filters gives the same or even better performance than learning 100 appearance filters, thus confirming that our approach reduces potential redundancy.

As expected, learning appearance filters using SC [8] improves performance over k-means (at a parity of dictionary size), but it is much slower: AUPRC are 0.8748 vs 0.8557 on IVCM, 0.77 vs 0.7460 on VC6, 0.7955 vs 0.7857 on BF2D; speed 25-30 mins vs a few seconds. However, learning 10 context filters with k-means yields better AUPRC than [8] (0.8866 vs 0.8748 on IVCM, 0.7911 vs 0.77 on VC6, 0.7966 vs 0.7955 on BF2D), although in [8] SC is used. Also, while learning with SC is time-consuming for filter banks larger than 100 (several days are reportedly needed to learn a filter bank of 121 filters), a few minutes are required to learn as many as 200 context filters with k-means, in case a larger dictionary is needed (e.g. for VC6). As a result, our method combining OOF with context filters learned using k-means outperforms significantly the method proposed in [8]: best AUPRC figures are 0.9078 vs 0.8748 on IVCM, 0.8368 vs 0.7700 on VC6, 0.8057 vs 0.7955 on BF2D.

4 Conclusion

Boosting hand-crafted features with learned filters has recently emerged as a successful technique to compensate for limits of HCFs and fully-learned approaches. We have proposed a novel combination method in which HCFs are paired with learned context filters to enhance pixel representation including inter-object relationships. Quantitative results suggest that our segmentation framework for curvilinear structures can be used for different image modalities and get top-rank performance running in a few minutes only. Our future work will investigate the combination of learned context filters with our recently proposed HCF,

SCIRD [15], and the design of new ones to exploit optimally the information obtained by the appearance model.

References

1. Ikram, M., Ong, Y., Cheung, C., Wong, T.: Retinal vascular caliber measurements: Clinical significance, current knowledge and future perspectives. Ophthalmologica (2013)
2. Annunziata, R., Kheirkhah, A., Aggarwal, S., Cavalcanti, B.M., Hamrah, P., Trucco, E.: Tortuosity classification of corneal nerves images using a multiple-scale-multiple-window approach. In: OMIA, MICCAI (2014)
3. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (eds.) MICCAI 1998. LNCS, vol. 1496, pp. 130–137. Springer, Heidelberg (1998)
4. Soares, J.V., Leandro, J.J., Cesar, R.M., Jelinek, H.F., Cree, M.J.: Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. IEEE TMI (2006)
5. Santamaría-Pang, A., Colbert, C.M., Saggau, P., Kakadiaris, I.A.: Automatic centerline extraction of irregular tubular structures using probability volumes from multiphoton imaging. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part II. LNCS, vol. 4792, pp. 486–494. Springer, Heidelberg (2007)
6. Law, M.W.K., Chung, A.C.S.: Three dimensional curvilinear structure detection using optimally oriented flux. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 368–382. Springer, Heidelberg (2008)
7. Honnorat, N., Vaillant, R., Paragios, N.: Graph-based geometric-iconic guide-wire tracking. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part I. LNCS, vol. 6891, pp. 9–16. Springer, Heidelberg (2011)
8. Rigamonti, R., Lepetit, V.: Accurate and efficient linear structure segmentation by leveraging ad hoc features with learned filters. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part I. LNCS, vol. 7510, pp. 189–197. Springer, Heidelberg (2012)
9. Becker, C., Rigamonti, R., Lepetit, V., Fua, P.: Supervised feature learning for curvilinear structure segmentation. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part I. LNCS, vol. 8149, pp. 526–533. Springer, Heidelberg (2013)
10. Hannink, J., Duits, R., Bekkers, E.: Crossing-preserving multi-scale vesselness. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part II. LNCS, vol. 8674, pp. 603–610. Springer, Heidelberg (2014)
11. Sironi, A., Lepetit, V., Fua, P.: Multiscale centerline detection by learning a scale-space distance transform. In: CVPR (2014)
12. Tu, Z., Bai, X.: Auto-context and its application to high-level vision tasks and 3d brain image segmentation. IEEE TPAMI (2010)
13. Coates, A., Ng, A.Y.: Learning feature representations with k-means. In: Neural Networks: Tricks of the Trade (2012)
14. Law, M.W.K., Chung, A.C.S.: An oriented flux symmetry based active contour model for three dimensional vessel segmentation. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 720–734. Springer, Heidelberg (2010)
15. Annunziata, R., Kheirkhah, A., Hamrah, P., Trucco, E.: Scale and curvature invariant ridge detector for tortuous and fragmented structures. In: A. Frangi et al. (eds.) MICCAI 2015, Part III. LNCS vol. 9351, pp. 588–595. Springer, Heidelberg (2015)