GLAND SEGMENTATION IN COLON HISTOLOGY IMAGES USING HAND-CRAFTED FEATURES AND CONVOLUTIONAL NEURAL NETWORKS

Wenqi Li[1⋆], *Siyamalan Manivannan*[1⋆], *Shazia Akbar*[†], *Jianguo Zhang*[⋆],
Emanuele Trucco[⋆], *Stephen J. McKenna*[⋆]
[⋆] CVIP, School of Science and Engineering, University of Dundee, Dundee, UK
[†]Center for Advanced Imaging Innovation and Research, NYU School of Medicine, New York, USA

ABSTRACT

We investigate glandular structure segmentation in colon histology images as a window-based classification problem. We compare and combine methods based on fine-tuned convolutional neural networks (CNN) and hand-crafted features with support vector machines (HC-SVM). On 85 images of H&E-stained tissue, we find that fine-tuned CNN outperforms HC-SVM in gland segmentation measured by pixel-wise Jaccard and Dice indices. For HC-SVM we further observe that training a second-level window classifier on the posterior probabilities – as an output refinement – can substantially improve the segmentation performance. The final performance of HC-SVM with refinement is comparable to that of CNN. Furthermore, we show that by combining and refining the posterior probability outputs of CNN and HC-SVM together, a further performance boost is obtained.

Index Terms— Histology image analysis, Convolutional neural network, Gland segmentation

1. INTRODUCTION

Analysis of gland structures is an important component of histopathological examinations. In this paper we address the challenging problem of gland segmentation in histology images. Fig. 1 shows some H&E-stained slices of colon biopsies with glands annotated by pathologists. Our aim is to obtain annotations similar to these automatically.

In previous work on this problem, glandular structures have often been modelled explicitly, mainly relying on the detection of nuclei and lumen. For instance, Sirinukunwattana et al. [1] modelled each gland as a polygon with vertices positioned at nuclei near the gland's perimeter. Polygon configurations were sampled using Markov chain Monte Carlo simulation. This method achieved good segmentation accuracy on healthy glands but was often less accurate on cancerous glands. An alternative approach (though not specifically for gland segmentation) avoids explicitly modelling tissue's geometric structure; this is exemplified by various deep learning methods that have shown promise for histology image analysis tasks. For example, Chang et al. [2] proposed to learn a series of dictionary elements using multi-layer unsupervised learning for tumor histopathology classification. Cireşan et al. [3] designed mitosis detection methods in breast cancer histology images with convolutional neural networks. Cruz-Roa et al. [4] learned features using convolutional auto-encoders for skin cancer detection. Notably, deep convolutional neural networks (CNN) have outperformed the previous state-of-the-art for several visual object segmentation [5] and biomedical image segmentation [6, 7] tasks. A CNN model usually consists of multiple layers of non-linear functions and a very large number of parameters. Hierarchical image feature representations can be learned from images by training CNNs discriminatively. However, it remains to be demonstrated whether CNNs can easily represent the visual structure of glands, and how gland segmentation performed using CNN features compares to the use of classification based on more traditional computer vision features.

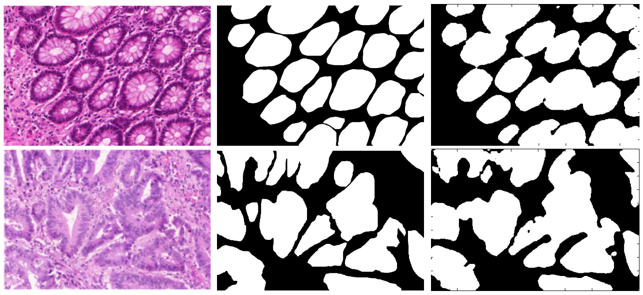

Fig. 1. Left column: examples of colon histology images. Middle column: gland annotations provided by pathologists. Right column: automatic segmentations obtained by fusing class posteriors from CNN and HC-SVM.

In this paper we tackle the gland segmentation problem with a window-based classification method. We conduct an evaluation of fine-tuned CNNs and state-of-the-art hand-crafted features with support vector machines (HC-SVM) [8]. Our contributions are three-fold. (1) We show that CNN outperforms HC-SVM in gland segmentation. (2) We show that training a second-level window classifier on the posterior probabilities can improve the segmentation performance of the HC-SVM method. The final performance of HC-SVM is then comparable to that of CNN. (3) We show that combining posterior probabilities output by CNN and HC-SVM using a second-level window classifier can further improve performance. The last column of Fig. 1 shows segmentation results produced by this method.

2. HISTOLOGY IMAGE FEATURES

Our window-based classification method starts with a classifier training phase. Firstly, image windows of size $W \times W$ ($W > 12$) are densely sampled from each image using a step size of 12 pixels in the horizontal and vertical directions. Features are then computed from each window using both a hand-crafted feature extractor and and fine-tuned CNNs. A binary classifier is trained to classify a window as either 'gland' or 'non-gland' based on extracted features. In the testing phase, given a test image, image windows are extracted

[1] These authors contributed equally to this work.

S. Manivannan is supported by EPSRC grant EP/M005976/1

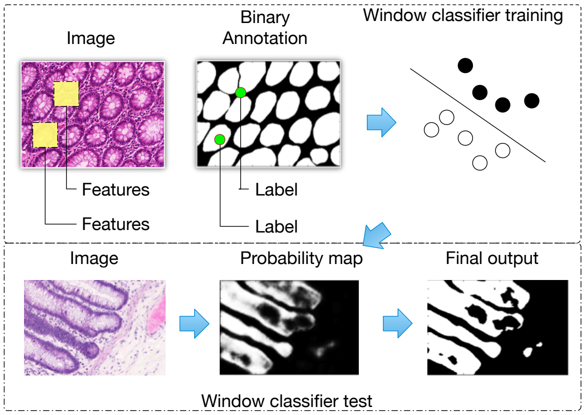

Fig. 2. A simplified overview of gland segmentation as window-based classification.

with a sliding window of size $W \times W$ shifted at a step size of 12 pixels, feature representations are computed, and the binary classifier applied to estimate the posterior probability of gland at the centre of each image window. These probabilities are thresholded to obtain the final gland segmentation (with a threshold set by using a grid search to maximise Dice indices on the training set). Fig. 2 illustrates this process of segmentation as window classification. In this section, we briefly summarise our image window feature representations.

2.1. Fine-tuned CNN features

We employed Alexnet [5] and Googlenet [9] deep network architectures. Both networks have shown remarkable image classification performance in the ImageNet large scale visual recognition challenge [10]. We utilised the weights pre-trained on ImageNet for both networks. In order to adapt the networks to our binary segmentation problem we fine-tuned the pre-trained weights. More specifically, (1) we replaced the last layer of each network (a 1000-way classifier designed for ImageNet classification) with a binary logistic regression; (2) we fed image windows densely sampled from histology images and gland/non-gland labels into the CNN training process to further update the pre-trained weights. We reduced the starting learning rates to $1/10$ of those used for ImageNet training.

For data augmentation during fine-tuning, additional image windows were obtained by rotation through angles of 90°, 180°, and 270°, and by randomly mirroring. The default window input dimensions are 227×227 and 224×224 for Alexnet and Googlenet respectively. We scaled[1] the histology image windows of size $W \times W$ (experiments with different W are reported in Section 3) to 256×256, and cropped a centred window to obtain windows of 227×227 pixels (224×224 for Googlenet). We used the Caffe library [11] for the fine-tuning process. For both networks we trained with $10,000$ iterations using an NVIDIA GPU[2]. In the test phase, given a test image window, the fine-tuned CNN model can directly output the posterior probability of gland. The computational time for testing one image (size: 522×775) was approximately $20s$.

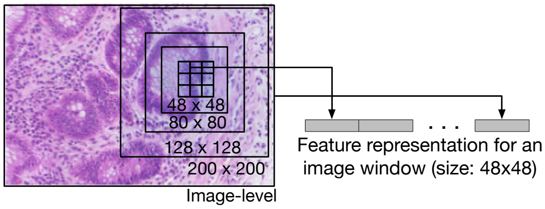

Fig. 3. Hand-crafted feature based window representation.

2.2. Hand-crafted features

The hand-crafted features in our method were designed to capture rich contextual information for window classification. The window size W was set to $W = 48$. Within each window, root-SIFT (SIFT) [12], vectorized raw-pixel values (RAW), and multi-resolution local patterns (mLP) [8] were extracted from patches of size 16×16 with a step size of 2 pixels. For each feature type, features extracted from the three color channels (R, G, and B) were concatenated. Locality-constrained Linear Coding (LLC [13]) with a dictionary size of D (experiments with different dictionary sizes were reported in Section 5) and sum pooling were applied to encode each window.

To capture more contextual information, we also computed and concatenated the window representation from concentric windows of size 80×80, 128×128, and 200×200, as well as the entire image (as shown in Fig. 3). In addition to this, to capture local structure information, the 48×48 window was divided into nine 16×16-pixel square regions and the feature representations obtained from these were also concatenated with the window representation. We used power and L2 normalizations [14] to normalize the pooled encoded features from each individual window/region. The final dimensionality of the window descriptor was 14 (regions) $\times D$ (size of the dictionary) $\times 3$ (features).

We trained four linear SVM classifiers [15], each using the rotated versions ($\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$) of the original training set. Platt scaling [16] was used to convert the SVM outputs to probabilities. During testing, the probabilities from these classifiers were averaged to obtain the final probability map for each test image. (For more information please refer to Ref. [8]).

3. FUSION OF HAND-CRAFTED AND CNN FEATURES

3.1. Fully-connected layer outputs as features

In Alexnet, the outputs of fully-connected layers can be treated as features [17]. We adopted the first and second fully-connected layer outputs (denoted as “FC_1” and “FC_2”) of the fine-tuned Alexnet as features. The dimensionalities of FC_1 and FC_2 are both 4096. We trained linear SVMs to classify FC_1 and FC_2 respectively, instead of using the final probability output of the fine-tuned CNN. We denote this approach as “FC_1_SVM” and “FC_2_SVM”.

3.2. Feature-level fusion

Feature-level fusion consisted of simply concatenating FC_1 or FC_2 with the hand-crafted features computed from the window centred at the same sampling point. A linear SVM was trained to classify the final features as gland or non-gland. We denote this approach as “Hand-crafted+FC_1” and “Hand-crafted+FC_2”.

[1] We used the `imresize` function in the MATLAB Image Processing Toolbox with the `bicubic` option.

[2] The Tesla K40 GPU used for this research was donated by the NVIDIA Corporation

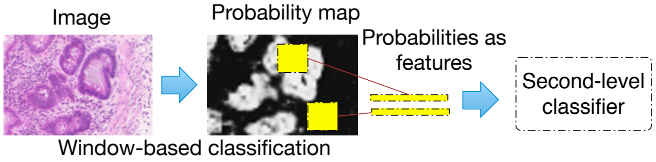

Fig. 4. Probability-level refinement.

3.3. Probabilitiy-level refinement and fusion

To refine the posterior probability output of each method, we used second-level window classifiers as illustrated in Fig. 4. Image windows were densely sampled from posterior probability maps and vectorised as the window representation.

We also considered probability-level fusion. Windows were densely sampled from CNN and HC-SVM probabilities and concatenated as the window representation. A second-level linear SVM was trained to classify the final representations.

To reduce the computational cost, we down-sampled each probability map by a factor of 5. Windows were densely sampled from probability maps with a step size of 1 in the horizontal and the vertical directions. The size of the windows was 5 × 5.

4. IMAGE DATA AND MEASUREMENTS

A subset[3] of the Warwick-QU dataset [1] consisting of 85 images of H&E-stained slides (37 benign and 48 malignant cases) together with annotations of glands by experienced pathologists was available. by Sirinukunwattana et al. [1]. The resolution of each image is 0.62 µm/pixel. Seventy-nine images are 522 × 775 pixels in size; the remaining six are 453 × 589 pixels. All results reported in the following sections were based on two-fold cross validation on this dataset. Mean values and standard deviations were calculated across all test images.

Following Sirinukunwattana et al. [1], we adopted pixel-wise Jaccard and Dice indices as segmentation performance metrics. Given a set of pixels marked as glandular structure ground truth, G, and a set of pixels segmented as glandular structures, O, both indices measure similarity between G and O. The Jaccard index is calculated as $\text{Jaccard}(G, O) = |G \cap O|/|G \cup O|$ and the Dice index is calculated as $\text{Dice}(G, O) = 2|G \cap O|/(|G| + |O|)$, where $|\cdot|$ denotes set cardinality. Both indices produce scores between 0 and 1, where 1 indicates perfect segmentation.

5. RESULTS

Effect of window size in CNN models

Table 1 reports segmentation results based on directly thresholding fine-tuned CNN outputs. Alexnet and Googlenet showed similar performance with both performing well at window size 32 × 32. We used the fine-tuned networks at window size 32 × 32 in the following comparison and fusion experiments.

Table 1. CNN results for different window sizes and networks.

Network	$W \times W$	Jaccard	Dice
Alexnet	32 × 32	**0.72 ± 0.13**	**0.83 ± 0.09**
Alexnet	64 × 64	0.68 ± 0.11	0.81 ± 0.09
Alexnet	96 × 96	0.63 ± 0.12	0.77 ± 0.10
Googlenet	32 × 32	0.71 ± 0.14	0.82 ± 0.11
Googlenet	64 × 64	0.68 ± 0.12	0.81 ± 0.10
Googlenet	96 × 96	0.64 ± 0.12	0.77 ± 0.10

Table 2. Segmentation results (Dice indices) for different dictionary sizes with individual hand-crafted features.

Type	100	200	500	1000
SIFT	0.74 ± 0.12	0.75 ± 0.11	0.74 ± 0.13	0.75 ± 0.11
RAW	0.74 ± 0.12	0.75 ± 0.10	0.76 ± 0.10	**0.77 ± 0.08**
mLP	0.76 ± 0.09	**0.77 ± 0.09**	0.76 ± 0.09	0.76 ± 0.08

Table 3. Effect of data augmentation for the hand-crafted features (the dictionary size was fixed to 200).

Type	without data augmentation		with data augmentation	
	Jaccard	Dice	Jaccard	Dice
ALL	0.65 ± 0.12	0.78 ± 0.09	**0.67 ± 0.12**	**0.80 ± 0.09**

Table 4. Comparison of features and their concatenation (all with an SVM classifier).

Type	Jaccard	Dice
FC_1_SVM	**0.72 ± 0.12**	**0.83 ± 0.09**
FC_2_SVM	**0.72 ± 0.12**	**0.83 ± 0.10**
Hand-crafted+FC_1	0.70 ± 0.11	0.82 ± 0.08
Hand-crafted+FC_2	0.71 ± 0.11	0.82 ± 0.08

Table 5. Results for probability-level refinement and fusion.

Type	Jaccard	Dice
Hand-crafted	0.71 ± 0.11	0.83 ± 0.09
Alexnet	0.73 ± 0.13	0.84 ± 0.10
Googlenet	0.72 ± 0.15	0.82 ± 0.12
Alexnet+Googlenet	0.74 ± 0.14	0.84 ± 0.10
Hand-crafted+Alexnet	**0.77 ± 0.12**	**0.87 ± 0.08**
Hand-crafted+Googlenet	0.75 ± 0.12	0.85 ± 0.09
Hand-crafted+Alexnet+Googlenet	**0.77 ± 0.11**	**0.87 ± 0.08**

Effect of dictionary size in hand-crafted features

Table 2 lists Dice indices for each individual hand-crafted feature at different dictionary sizes. Changing the size of the dictionary did not considerably affect the segmentation performance. In Table 2 RAW and mLP gave better segmentation results than SIFT. However, when combining the representations obtained from all these features ('ALL') the performance was further improved (Table 3). Table 3 investigates the effect of data augmentation when all the features were used. Training with rotated image windows considerably improves the overall segmentation. Therefore, in the following experiments all the features were used with data augmentation, and the dictionary size was set to 200.

Fusion of hand-crafted and CNN features

Table 4 reports Jaccard and Dice indices of segmentations obtained using fully-connected layer outputs as features (Section 3.1), as well as using feature-level fusion (Section 3.2). Table 5 shows the segmentation results obtained by applying refinement or fusion classifiers to different probability maps (Section 3.3). Fig. 5 visualises two types of errors for each method.

[3] This subset was released as part of the Gland Segmentation (GlaS) challenge contest held in conjunction with MICCAI 2015. `http://www2.warwick.ac.uk/fac/sci/dcs/research/combi/research/bic/glascontest/`

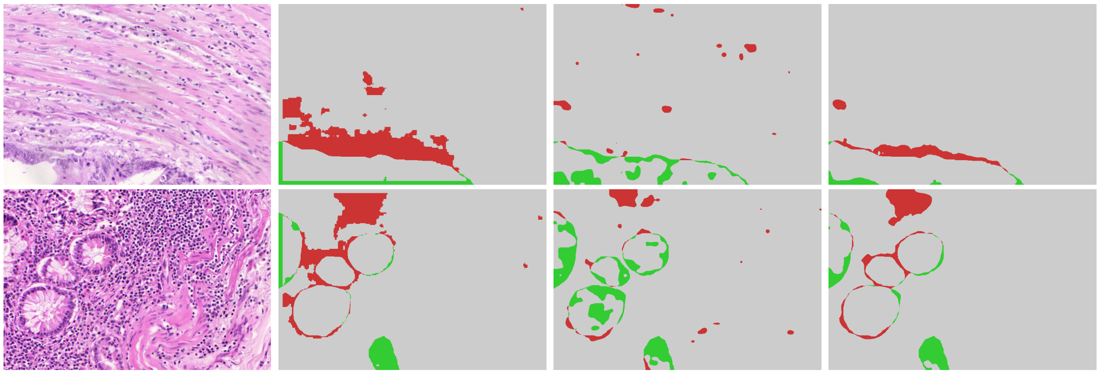

Fig. 5. Visualisations of segmentation errors. Columns from left to right correspond to original image, results using HC-SVM, results using Alexnet features, and results using hand-crafted+Alexnet with probability-level fusion. Green indicates false negatives; red indicates false positives; grey indicates correct predictions.

6. DISCUSSION

Experiments showed that CNN features outperformed a combination of three hand-crafted window representations. Interestingly, feature-level fusion of hand-crafted and CNN features did not give better performance than using CNN alone. The results in Table 1 and Table 4 also indicate that the fine-tuned CNN models can readily give good binary predictions (Table 1 first row); training a separate binary classifier on the fully-connected layer outputs is not necessary. At probability-level, refinement of HC-SVM provides an improvement over the HC-SVM method. Similarly refining or fusing Alexnet or Googlenet showed no segmentation improvement. However, fusing CNN and HC-SVM did result in a further improvement, achieving the best segmentation of all the methods we evaluated. This indicates that these two contrasting approaches make different errors and are complementary.

7. REFERENCES

[1] K. Sirinukunwattana, D. R. J. Snead, and N. M. Rajpoot, "A stochastic polygons model for glandular structures in colon histology images.," *IEEE Transactions on Medical Imaging*, vol. 34, no. 11, pp. 2366–2378, 2015.

[2] Hang Chang, Yin Zhou, Alexander Borowsky, Kenneth Barner, Paul Spellman, and Bahram Parvin, "Stacked predictive sparse decomposition for classification of histology sections," *IJCV*, vol. 113, no. 1, pp. 3–18, 2015.

[3] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *MICCAI*, vol. 8150 of *LNCS*, pp. 411–418. Springer, 2013.

[4] Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto González Osorio, "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection," in *MICCAI*, vol. 8150 of *LNCS*, pp. 403–410. Springer, 2013.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, vol. 9351 of *LNCS*, pp. 234–241. Springer, 2015.

[7] Yuanpu Xie, Xiangfei Kong, Fuyong Xing, Fujun Liu, Hai Su, and Lin Yang, "Deep voting: A robust approach toward nucleus localization in microscopy images," in *MICCAI*, vol. 9351 of *LNCS*, pp. 374–382. Springer, 2015.

[8] Siyamalan Manivannan, Wenqi Li, Shazia Akbar, Ruixuan Wang, Jianguo Zhang, and Stephen J McKenna, "An automated pattern recognition system for classifying indirect immunofluorescence images of HEp-2 cells and specimens," *Pattern Recognition*, vol. 51, pp. 12–26, 2016.

[9] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *IJCV*, pp. 1–42, 2014.

[11] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM MM*, 2014, pp. 675–678.

[12] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*, 2012, pp. 2911–2918.

[13] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010, pp. 3360–3367.

[14] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the Fisher kernel for large-scale image classification," in *ECCV*, 2010, pp. 143–156.

[15] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[16] John Platt et al., "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[17] Ali S Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *CVPR Workshops, 2014 IEEE Conference on*, 2014, pp. 512–519.