Cell Segmentation in Digitized Pap Smear Images Using an Ensemble of Fully Convolutional Networks

G. Bogacsovics, A. Hajdu and B. Harangi

Faculty of Informatics, University of Debrecen, Hungary
{bogacsovics.gergo, hajdu.andras, harangi.balazs}@inf.unideb.hu

Abstract— This paper presents a method that provides reliable performance regarding cell segmentation in digitized Pap smear images. Since our final goal is the early detection of cervical cancer using scanned smear images, the proper segmentation of cells is of utmost importance. Our approach uses segmentation predictions from fully convolutional networks (FCNs) in addition to the original scanned image as its input. Our method transforms these input images to a final segmentation using a dedicated FCN architecture. Thus, our approach can be considered an ensemble-based one and outperforms state-of-the-art segmentation algorithms, achieving close to 93% accuracy and a Dice score of more than 69%.

I. Introduction

In this paper, we present a part of an automatic screening system that is dedicated to the field of cytology. This part of the system deals with the reliable segmentation of cells that are present in the high-resolution, digitized Pap smear images. During its development, we have considered state-of-the-art machine learning methodologies to reach high segmentation accuracy. Our final goal is the creation of a fully automated system that can recognize cancerous cells in digitized cytological specimens with sufficiently high reliability and is suitable for clinical applications. The final software will be able to automatically rank the smear images, thus allowing the practicing clinicians to always focus on patients with the most severe conditions, making the treatment process faster and more efficient. By using a similar automatic screening system, the governments could save money and human efforts next to the developments of its outpatient services.

During the development of the system, our intention is to maximize the accuracies of the sub-components to reach the highest overall diagnostic accuracy which is expected to be close to that of a human grader. Using this computer-aided diagnosis (CAD) system, we plan to rank the specimens according to their severity levels and forward them to secondary grading to decrease the burden of the graders. In this way, we can ensure that the most diseased digitalized cytological specimens will be investigated by human graders, too. Thus, the usage of our CAD system could improve public health services.

At the beginning of the 21st century, a few automated screening systems are available with limited applicability in this field. They are based on computer image analysis techniques for screening and try to exclude the surely negative samples from the consequent investigation procedure to decrease the number of specimens. To the best of our knowledge, the most widely used automatic solutions as of now are the Hologic ThinPrep Imaging System [1] and the Focal Point Slide Profiler [2]. These have been approved by the FDA and operate in high-volume reference laboratories under human supervision. Unfortunately, unlike our system, the Hologic ThinPrep Imaging System can only analyze ThinPrep Pap Test slides which have much higher costs than the most commonly applied Papanicolaou smear test [3]. The other solution, the Focal Point Slide Profiler also has some notable drawbacks, as it can eliminate only up to 25% of the lowest-risk slides to allow the cytologists to focus on the highest-risk slides. Our automatic system could overcome the limitations of these two solutions, as it could process the most commonly applied Pap smear test images and could also rank the slides by the level of risk more accurately. The official procedure of taking the Papanicolaou smear test begins by opening the vaginal canal with a speculum and collecting cells at the outer opening of the cervix. After that, the collected cells are fixed on the glass slide and this specimen is placed into a large capacity whole slide scanner which can digitize it by generating a digital color image with a resolution of $100,000 \times 220,000$ pixels. The resulted image can contain more than 10,000 cells at a 40x magnification as it can be seen in Figure 1.

Our automatic screening system aims to localize and segment each cell from this high-resolution image with high sensitivity and specificity. The workflow of this type of system can be formed by the following steps: the proper segmentation of the individual cells; a pre-trained deep learning-based algorithm that classifies all of the segmented cells as healthy or unhealthy. In the case that any of the cells is considered pathologically diseased, the whole test will be investigated by a cytologist, as well. The high sensitivity is crucial because of the mortality of the Human Papillomavirus (HPV) and other cervical cancers. Based on the statistics, $2-3$ million abnormal Pap smear results are found each year [4] in the United States.

To ensure high sensitivity for the cell classification, we should find all the cells from the specimen, while excluding cell debris, blood platelets, and any other impurities before the additional processing steps. For

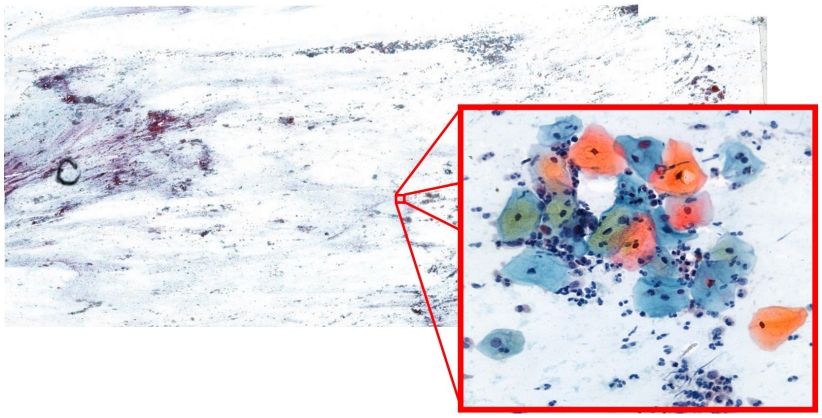

Figure 1. A sample scanned image about cytological specimen, where the red box shows cells with 40x magnification

this aim, we have developed a fusion-based solution to find and segment each cell with its nucleus and plasma with high accuracy. In our work, we primarily focus on fully convolutional networks (FCNs) [5], which as their name suggests only use convolutional layers to get a segmentation mask from an input image. We build our proposed method upon them, as they have been widely used for the segmentation of Pap smear images with great success [6] [7]. Namely, we consider the segmentation predictions of some fully convolutional networks besides the original scanned image as its input. Our method transforms this input into a final segmentation result with a dedicated FCN architecture trained accordingly. Thus, our approach can be considered as an ensemble-based one that uses some segmentation outputs to improve and support the final segmentation result. In the following sections, we introduce the applied fully convolutional networks and show how to combine their output to reach a higher accuracy than any of its members with this region-based ensemble. We also compare our results with our previous work [8], where we combined the outputs of different algorithms by using simple majority voting. We compare these results with some state-of-the-art solutions as well, like the U-Net [9] and GSCNN [10] networks. The U-Net network is based on fully convolutional layers as well, but uses a unique architecture that has a contracting and expanding part, which can help with localization. GSCNN on the other hand uses two separate paths, known as streams to process the input image. This way, the network can process the classic features and the shapes on parallel streams and then uses gates to combine these. We evaluate both our previous approach and these other state-of-the-art networks on the dataset introduced in IV and compare them with our new method to show how our method surpasses both our previous results and the state-of-the-art networks.

The rest of the paper is organized as follows. In Section 2, we describe our methodology to create an ensemble for segmentation purposes. Our experimental results including the descriptions of the data sets are enclosed in Section 3, while some conclusions are drawn in Section 4.

II. The Proposed Segmentation Algorithm

In this section, we introduce our fusion-based fully convolutional network for cell segmentation, which instead of taking solely the image as the input, receives both the original three-channel (RGB) image and the outputs of other FCN [5] algorithms. The reason behind this lies in our observation that there are many cases when even though the different FCN (FCN-8, FCN-16 and FCN-32) algorithms provided very different and disjoint outputs, these outputs could still be aggregated in such a way that the combined result would have had a higher accuracy. Such a common scenario occurred when the various algorithms found their distinct group of cells on a given smear image. This phenomenon can be observed in Figure 2.

It can easily be seen that a standard aggregation model (e.g. majority voting, statistical combination) would have problems in cases where each algorithm finds different parts of the cells or cell groups. So we aim to provide an efficient solution for the combination of the segmentation outputs, or in other words, we propose a region-based ensemble system. Consequently, our method used the information gathered by the individual segmentation algorithms, while still being able to make its own decision by involving the original image as well.

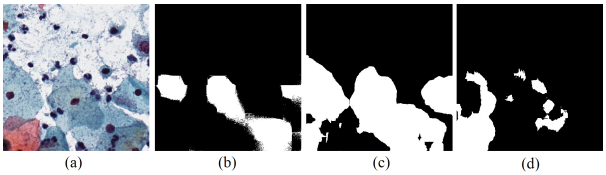

Figure 2. A sample input image (a) and the outputs of the FCN-32 (b), FCN-16 (c) and FCN-8 (d) algorithms

To achieve this, first, we train some FCN algorithms, namely the FCN-8, FCN-16 and FCN-32 networks. As pointed out in [5], these algorithms differ in their main architectures and the amount of upsampling they use. The FCN-8 is based on AlexNet [11], while the other two use the architecture of the VGG-16 [12] network. The numbers in the names of these models represent the amount of upsampling used for each respective network.

After training the FCN networks, we combine their outputs. To avoid using only their (sometimes) improper segmentation results, we also use the original image as input. In this way, the final ensemble model could consider the original input image and the segmentation results together and could learn how it should combine them to reach the most accurate output.

The member algorithms assign values $FCN_i(p_{x,y}) \in [0,1]$ (where (i=8,16,32)) to each $p_{x,y}$ pixel of the input image I to indicate their confidence whether $p_{x,y}$ belongs to a cell (plasm or nucleus) or not. Instead of using pixel-wise combination after thresholding, like an element-wise multiplication or majority voting, we consider region-based combination. Our idea is to concatenate the input image and the outputs of the members into a joint matrix $I_{con}(x,y) = [I(x,y); FCN_8(p_{x,y}); FCN_{16}(p_{x,y}); FCN_{32}(p_{x,y})]$, where $I(x,y)$ is a 3D vector that contains the normalized intensity values of the original input image at (x,y) regarding the red, green and blue channels. I_{con} is provided as an input to the $FCN_{Comb}(p_{x,y}) \rightarrow \{0,1\}$ (see Figure 3) which results in the required region-based combination by applying the appropriate convolutional filters with weights found during the second stage of the training.

This was achieved by implementing FCN_{Comb} as a regular FCN-32 architecture, where we increased the number of input channels. Thus both the outputs of the FCN algorithms and the input image can pass through each convolutional and deconvolutional layer at the same time, ensuring region-based combination by using convolutional operators instead of pixel-wise ones.

It can also be seen how our algorithm can receive the outputs of multiple pre-trained models with the original input image to determine the final segmentation output. In our work, we showcase the results of several actual implementations of the proposed algorithm. One difference between these is the number of pre-trained models that they use. Our reasoning for trying out multiple variants is that we wanted to experiment with minimizing the amount of extra information that is being given to the model and how this reduction affects performance.

III. Hyperparameters

We systematically searched for the optimal hyperparameters for each algorithm: both the baselines and the proposed one. During this process, we explored a number of different combinations and ranges by running several manual experiments and noted the best performing variations.
For the batch size, we could only use a maximum of 4 due to hardware limitations as we carrioud out the experiments on a computer with a GTX 1070 graphics card. In the case of our proposed model, we found that using lower learning rates enabled us to reduce the oscillation of the learning loss and make the learning more stable. For learning rate, the oscillation became noticably smaller under 0.001, and in the end, 0.0001 produced the best results, which we used to train the proposed models. To compensate for this low learning rate, we had to increase the number of epochs. We found that 200 epochs worked best for our experiments. For the other parameters, such as stride, padding, etc. we mostly used the ones recommended by [5]. The parameters of the first convolutional layer were however changed to compensate for the bigger input images, as can be seen in Figure 3.

IV. Dataset and Experimental Results

In this section, we present the measured individual performances of the traditional FCN algorithms and that of the actual implementations of our proposed ensemble-based solution.

IV-A. Dataset

Our dataset contains digital images derived from scanned results of the Pap smear tests. Each sample has been manually annotated by clinical experts. Every entry in this dataset is thus made up of two elements: the scanned image and the manual segmentation corresponding to it. The set of data is divided into three parts: we trained the FCN algorithms on the first part, the combined network on the second one, and evaluated their respective performances on the third part. The three parts consist of 1284, 416 and 557 images, respectively, of size 500x500 pixels. This way, we have avoided excessive training of the combined network on the same data that the individual FCN algorithms were trained on.

IV-B. Quantitative results

We use the abbreviation C to refer to our combined network with some suffixes, which denote the extra inputs used. For example C_{16-8} means that the combined network receives the raw input image, as well as the outputs of the trained FCN-16 and FCN-8 algorithms.

To evaluate the trained networks, we used the indicators true positive (TP), false positive (FP), true negative

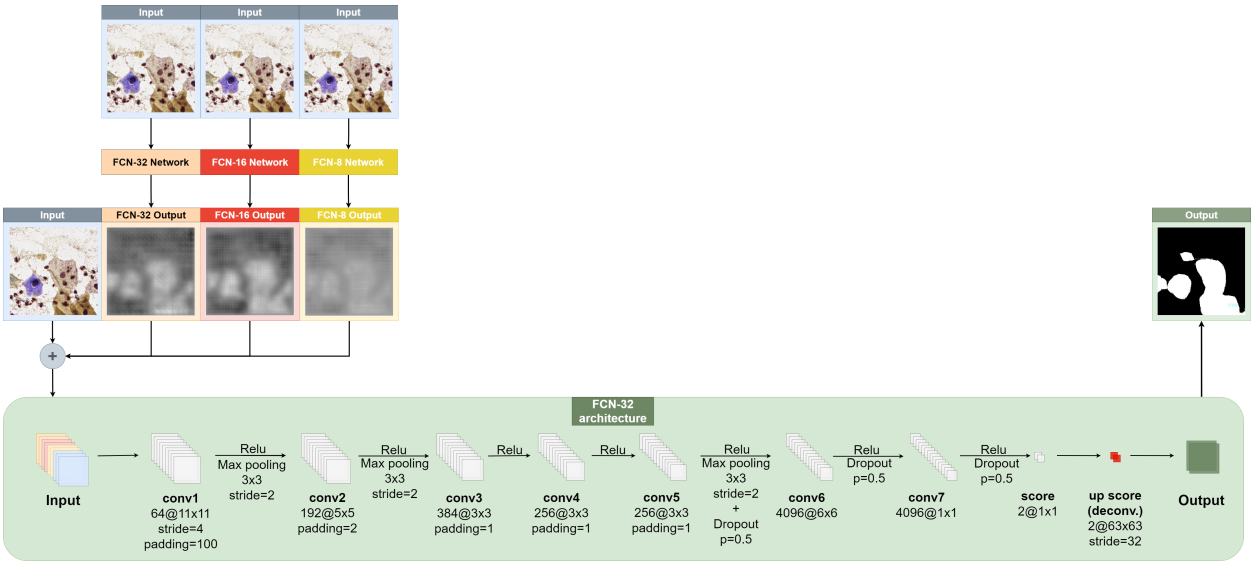

Figure 3. An illustrative overview of the proposed algorithm with genuine inputs and outputs. The system receives both the outputs of the FCN algorithms and the original RGB image as input and produces a segmented image as its output.

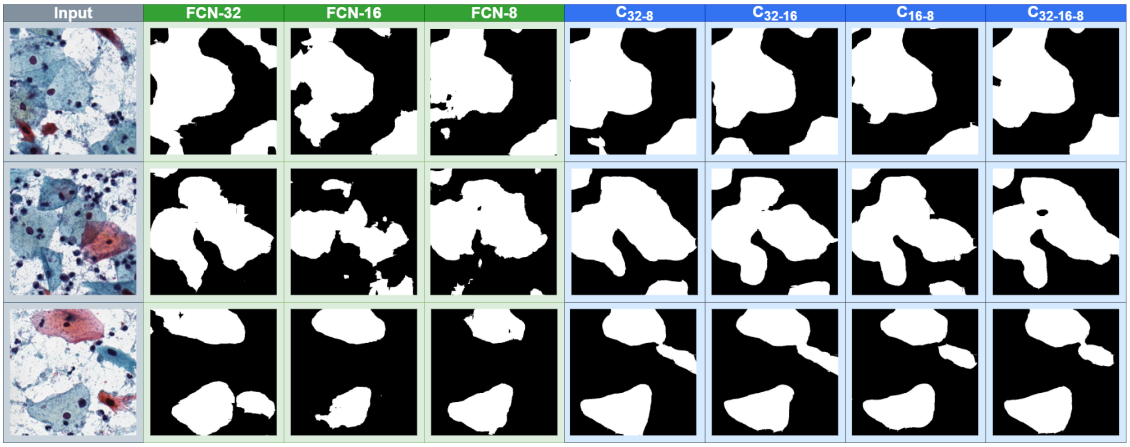

Figure 4. A comparison between the performances of the individual FCN algorithms and the combined architectures

(TN) and false negative (FN). TP means the number of pixels that have been labeled as part of the cell plasma both in the prediction and the ground truth images. TN is similar to TP, but it is based on the pixels that belong to the background. The FP and FN counts represent the pixels that have been incorrectly labeled as negative or positive throughout the prediction, respectively. Based on these indicators, we have calculated the following measures for each algorithm: accuracy (ACC), intersection over union (IoU) and dice score (DSC). Moreover, we have used cross-validation, during which we shuffled the previously mentioned three parts of the data around so that we could evaluate the performance of the networks on different test sets. Table 1 shows the overall results at 95% confidence level.

It can be seen that the combined networks yielded better results than any of the member FCN architectures (e.g. a +5% improvement compared to even the best performing FCN algorithm (FCN-8) for IoU and +4% for DSC) which were used originally as a cell segmentation algorithm in [15] and other state-of-the-art re-implemented methods based on [13] and [14] for the

Table 1. Results on the test dataset

Algorithm	*ACC*	*IoU*	*DSC*
FCN-32 [5]	0.915 ± 0.054	0.497 ± 0.161	0.660 ± 0.144
FCN-16 [5]	0.913 ± 0.063	0.503 ± 0.143	0.666 ± 0.127
FCN-8 [5]	0.919 ± 0.037	0.507 ± 0.180	0.668 ± 0.158
sota [13]	0.775	0.343	
Ens_1 [8]	0.923 ± 0.022	0.534 ± 0.239	0.688 ± 0.205
Ens_2 [8]	0.923 ± 0.020	0.534 ± 0.243	0.688 ± 0.208
DeepLabv3[14]	0.889 ± 0.117	0.487 ± 0.039	0.655 ± 0.035
U-Net [9]	0.917 ± 0.063	0.504 ± 0.224	0.662 ± 0.199
GSCNN [10]	0.909 ± 0.091	0.514 ± 0.185	0.674 ± 0.162
C_{32-8}	0.926 ± 0.034	0.530 ± 0.195	0.688 ± 0.168
C_{32-16}	0.928 ± 0.036	0.534 ± 0.191	0.691 ± 0.163
C_{16-8}	**0.928 ± 0.031**	**0.537 ± 0.203**	**0.693 ± 0.173**
$C_{32-16-8}$	0.927 ± 0.040	0.531 ± 0.175	0.687 ± 0.147

most important metrics of segmentation, namely *IoU* and *DSC*. The reason for our focus on these two metrics is that the manually annotated dataset used to train the algorithms contained some imperfect annotations, meaning that in some cases the outlines of the cells were bigger and rougher than needed. Consequently, there were cases when the algorithms produced even more accurate masks than the ground truth, resulting in a higher number of FNs instead of TNs when comparing them. This phenomenon motivated us to focus on measures that do not take into account the TNs. It is also imperative to note that as our test set contained 557 images with widths and heights of 500-500, the total number of pixels in the test set was 139,250,000. By this logic, e.g. in terms of accuracy even a 1% improvement leads to an increase of 1,392,500 correctly labeled pixels. It is also important to note that the accuracies of the baseline models were all well over 90%, making these improvements even more substantial and valuable (since in this range even tiny improvements require a lot of effort, engineering and extra computation). Some comparisons regarding the outputs of the standard FCN models and our proposed architecture can be seen in Figure 4.

V. Conclusions

We have proposed a cell segmentation approach to combine multiple trained fully convolutional networks to obtain a model that exceeds the performances of all of these individual models. We worked on the problem of segmentation of cells on digitized Pap smear images, which is a complex issue, and compared the results of the traditional FCN algorithms with actual implementations of our proposed method. We have shown that any combination proved to be more accurate than any of these traditional algorithms and yielded better segmentation results.

Moreover, we have also noted that giving both the outputs of these FCN algorithms and the input to the combined networks can be a usable solution to achieve higher segmentation accuracy. Namely, this way not only can the model combine the different outputs, but it also comes up with its own decisions on how to combine them and what to do with the different segments of these outputs (e.g. link them assuming that cells are connecting them), thus further improving the preciseness of the model. This improvement could be seen from the increase in the number of cells found and from the more accurate extraction of the cells when compared to the traditional methods. In the future, we also plan to extend the proposed framework by gathering additional data and re-training the networks, as well as evaluating more architectures as the base of our proposed model.

Acknowledgements

Research reported in this publication was supported by the ÚNKP-21-3-I-DE-99 and the ÚNKP-20-5-DE-31 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund. Research was also supported in part by the Janos Bolyai Research Scholarship of the Hungarian Academy of Sciences and the GINOP-2.2.1-18-2018-00012 supported by the European Union, co-financed by the European Social Fund.

References

[1] C. V. Biscotti, A. E. Dawson, B. Dziura, L. Galup, T. Darragh, A. Rahemtulla, and L. Wills-Frank, "Assisted primary screening using the automated thinprep imaging system," *American journal of clinical pathology*, vol. 123, no. 2, pp. 281–287, 2005.

[2] D. Chute, H. Lim, and C. S. Kong, "BD focalpoint slide profiler performance with atypical glandular cells on surepath papanicolaou smears." *Cancer cytopathology*, vol. 118, no. 2, pp. 68–74, 2010.

[3] G. N. Papanicolaou and H. F. Traut, "The diagnostic value of vaginal smears in carcinoma of the uterus," *Am. J. of Obstet. and Gynec.*, vol. 42, no. 2, pp. 193 – 206, 1941.

[4] R. P. Insinga, A. G. Glass, and B. B. Rush, "Diagnoses and outcomes in cervical cancer screening: A population-based study," *Am. J. of Obstet. and Gynec.*, vol. 191, no. 1, pp. 105 – 113, 2004.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[6] E. Hussain, L. B. Mahanta, C. R. Das, M. Choudhury, and M. Chowdhury, "A shape context fully convolutional neural network for segmentation and classification of cervical nuclei in pap smear images," *Artificial Intelligence in Medicine*, vol. 107, p. 101897, 2020.

[7] L. Zhang, M. Sonka, L. Lu, R. M. Summers, and J. Yao, "Combining fully convolutional networks and graph-based approach for automated segmentation of cervical cell nuclei," *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*. IEEE, 2017, pp. 406–409.

[8] B. Harangi, J. Toth, G. Bogacsovics, D. Kupas, L. Kovacs, and A. Hajdu, "Cell detection on digitized pap smear images using ensemble of conventional image processing and deep learning techniques," *11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pp. 38–42, 2019.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[10] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-scnn: Gated shape cnns for semantic segmentation," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5229–5238.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] Z. Lu, G. Carneiro, and A. P. Bradley, "Automated nucleus and cytoplasm segmentation of overlapping cervical cells," *Interna-*

tional Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2013, pp. 452–460.

[14] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[15] P. Naylor, M. Laé, F. Reyal, and T. Walter, "Nuclei segmentation in histopathology images using deep neural networks," *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, April 2017, pp. 933–936.