VOCA: Cell Nuclei Detection In Histopathology Images By Vector Oriented Confidence Accumulation

Chensu Xie[1,2] XIC3001@MED.CORNELL.EDU
Chad M. Vanderbilt[2] VANDERBC@MSKCC.ORG
Anne Grabenstetter[2] GRABENSA@MSKCC.ORG
Thomas J. Fuchs[1,2] FUCHST@MSKCC.ORG
[1] *Weill Cornell Graduate School of Medical Sciences, Cornell University, New York, USA*
[2] *Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, USA*

Abstract

Cell nuclei detection is the basis for many tasks in Computational Pathology ranging from cancer diagnosis to survival analysis. It is a challenging task due to the significant inter/intra-class variation of cellular morphology. The problem is aggravated by the need for additional accurate localization of the nuclei for downstream applications. Most of the existing methods regress the probability of each pixel being a nuclei centroid, while relying on post-processing to implicitly infer the rough location of nuclei centers. To solve this problem we propose a novel multi-task learning framework called vector oriented confidence accumulation (VOCA) based on deep convolutional encoder-decoder. The model learns a confidence score, localization vector and weight of contribution for each pixel. The three tasks are trained concurrently and the confidence of pixels are accumulated according to the localization vectors in detection stage to generate a sparse map that describes accurate and precise cell locations. A detailed comparison to the state-of-the-art based on a publicly available colorectal cancer dataset showed superior detection performance and significantly higher localization accuracy.

1. Introduction

Object detection in natural images has been defined as fitting tight bounding boxes around recognized objects. The best examples are the prevailing Fast/Faster-RCNN models (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015) and closely related techniques (Redmon et al., 2016; Liu et al., 2016; He et al., 2017). Cell nuclei detection on histopathology slides requires identification of millions of densely packed small objects per image. This is in contrast to these earlier deep learning works in which usually a few dominant objects are annotated. Due to the several orders of magnitude increase in numbers of objects detected per image, the performance of region proposal based detectors is sub-optimal on cell detection in histology images (Jeong et al., 2017). Further, obtaining annotation of thousands of nuclei bounding boxes is impractical due to the common case of weak nuclei boundaries and high workload of pathologists. To this end, these problems are usually formulated as predicting the (x, y) coordinates of the objects' center supervised by point labels (Fuchs et al., 2009).

Most deep learning approaches to cell nuclei detection are based on convolutional neural networks that predict the probability of each pixel being a nucleus centroid (Cireşan et al., 2013; Wang et al.,

2014; Xie et al., 2015b; Chen and Srinivas, 2016; Sirinukunwattana et al., 2016; Zhou et al., 2017; Raza et al., 2018). The final detection of the objects is achieved by identifying the peaks in the probability map using mean shift (Fuchs et al., 2009) or non-maximum suppression (Neubeck and Van Gool, 2006). Fast auto-encoded regression has recently been employed as a technique to explore improved speed and scalability in cell detection over the traditional sliding-window system (Xie et al., 2015a; Chen and Srinivas, 2016; Zhou et al., 2017). Current methods are designed to recognize the cell nuclei and rely on post-processing and *ad hoc* fine-tuning to implicitly infer cell locations, which leads to accumulation of localization error as the number of detected objects gets larger. We must emphasize that while the challenging cell detection is not a clinically useful end as a standalone task, the accurate coordinates of cell nuclei are simply the prerequisite for many downstream applications (e.g. multi-class cell detection for tumor micro-environment analysis, tumor architecture, etc).

To solve this problem, we propose a novel multi-task deep learning method for cell detection. Based on convolutional encoder-decoder, the model concurrently learns 1) binary confidence score, 2) localization vector and 3) weight of contribution for each pixel. In detection stage, the confidence scores are weighted and accumulated to the positions pointed by the localization vectors. We call this method vector oriented confidence accumulation (VOCA). We demonstrate that the three closely correlated but distinct tasks are mutually beneficial when trained as an integrated model (Section 5.1). VOCA explicitly learns the location of nuclei centroid and thus produces profoundly peaked accumulator maps which describe accurate and precise nuclei locations, and enables fast and robust post-processing (Section 5.2). Comparison experiments based on a publicly available colorectal cancer dataset (Sirinukunwattana et al., 2016) shows that our proposed method outperforms the existing methods in terms of F1 score for cell detection, and gives significantly higher nuclei localization accuracy (Section 5.3).

2. Related work

Early attempts at cell nuclei detection utilized human expert-designed features describing intensity distribution and morphological patterns (Cosatto et al., 2008; Al-Kofahi et al., 2010; Kuse et al., 2011; Arteta et al., 2012; Ali and Madabhushi, 2012; Veta et al., 2013; Vink et al., 2013). It is notable that many of these works confabulate the related but separate concepts of nuclei detection and segmentation. This confusion is likely because hand-crafted features are often shape oriented. These approaches tend to be brittle due to the significant heterogeneity of histology slides and cellular morphology and require additional engineering and tuning between different datasets.

Recent works employing deep learning for cell nuclei detection have achieved state-of-the-art results. Cireşan et al. (2013) utilized deep neural network to differentiate between mitotic nuclei and background. Cruz-Roa et al. (2013) and Xu et al. (2016) learned unsupervised features via auto-encoders for cell detection, which was extended by Wang et al. (2014) by combining hand-crafted features with deep learning. While object detection at its heart is the combination of object recognition and localization, these works depending on pixel-wise binary classification only considered the first task. Xie et al. (2015b) proposed a structured regression approach to predict the probability of each position being a nucleus centroid. Their regression targets embedded the localization information by formulating the score as a function of the distance (d) between each pixel and the nearest

ground truth nucleus. This spirit of integrating the two tasks was also followed by many other works. For example, Chen and Srinivas (2016) labeled pixels for lymphocytes detection by thresholding d. Sirinukunwattana et al. (2016) proposed a spatially constrained CNN (SC-CNN) regressing to a similar map and published a dataset for nuclei detection on colorectal cancer images. Zhou et al. (2017) developed a sibling fully convolutional network (FCN) architecture for simultaneous cell detection and fine-grained classification. Raza et al. (2018) proposed a framework to deconvolve filter-mapped CNN output for cell detection on lung cancer slides. Considering the variation in nuclei size, Koohababni et al. (2018) formulated each nucleus as a Gaussian peak with a maximum value on its centroid, and directly regress the means and standard deviations with a small image patch as input. Tofighi et al. (2018) utilized additional annotation to combine shape priors with deep features for cell detection. Notably, Ahmad et al. (2018) learned features by correlation filters and achieved state-of-the-art performance for nuclei detection on the previously mentioned colorectal dataset (Sirinukunwattana et al., 2016) against which several of the above mentioned works were benchmarked. In contrast to these works, VOCA formulates the cell nuclei detection problem as a multi-task approach, which disentangles rather than integrates the objectives, hypothesizing that simpler objectives can potentially improve model training and understanding.

3. Method

3.1. Deep multi-task learning

We propose a novel CNN based deep multi-task learning method for cell detection. Each pixel of a training image is scored with 3 tasks. Let $p_I[i,j]$ be the pixel at coordinate (i,j) of input image I, and $c_I[u,v]$ be the nearest ground truth annotation for a cell nuclei which is at position (u,v). $Conf_I$, Loc_I, and Wt_I be the target maps of confidence score, localization vector and weight of contribution of image I respectively. First,

$$Conf_I[i,j] = \begin{cases} 1, & \text{if } \|(u-i, v-j)\|_2 < r \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

r is the hyperparameter thresholding the proximity of cells. The confidence score target map indicates whether each pixel should be regarded as a nucleus. The second task

$$Loc_I[i,j] = (u-i, v-j), \quad \text{if } Conf_I[i,j] = 1 \tag{2}$$

is a vector describing the direction and magnitude that $p_I(i,j)$ needs to move to the location of its assigned ground truth $c_I(u,v)$. Note that only pixels labeled as foreground by the confidence map ($Conf_I[i,j] = 1$) are trained with this task. The third task scores $p_I[i,j]$ as:

$$Wt_I[i,j] = \sum_{c_I[u',v']} \mathbb{1}_{\|(u'-i,v'-j)\|_2 < r}(c_I[u',v']) \tag{3}$$

where $\mathbb{1}_{||(u'-i,v'-j)||_2<r}(c_I[u',v'])$ is an indicator function of whether a ground truth cell nucleus $c_I[u',v']$ is within euclidean distance r to $p_I[i,j]$. This task counts the number of cell nuclei that intersect at $p_I[i,j]$. Since the pixels lying in the intersection of cells are shared in confidence accumulation (cf. Section 3.3), their contribution should be up-weighted accordingly by Wt.

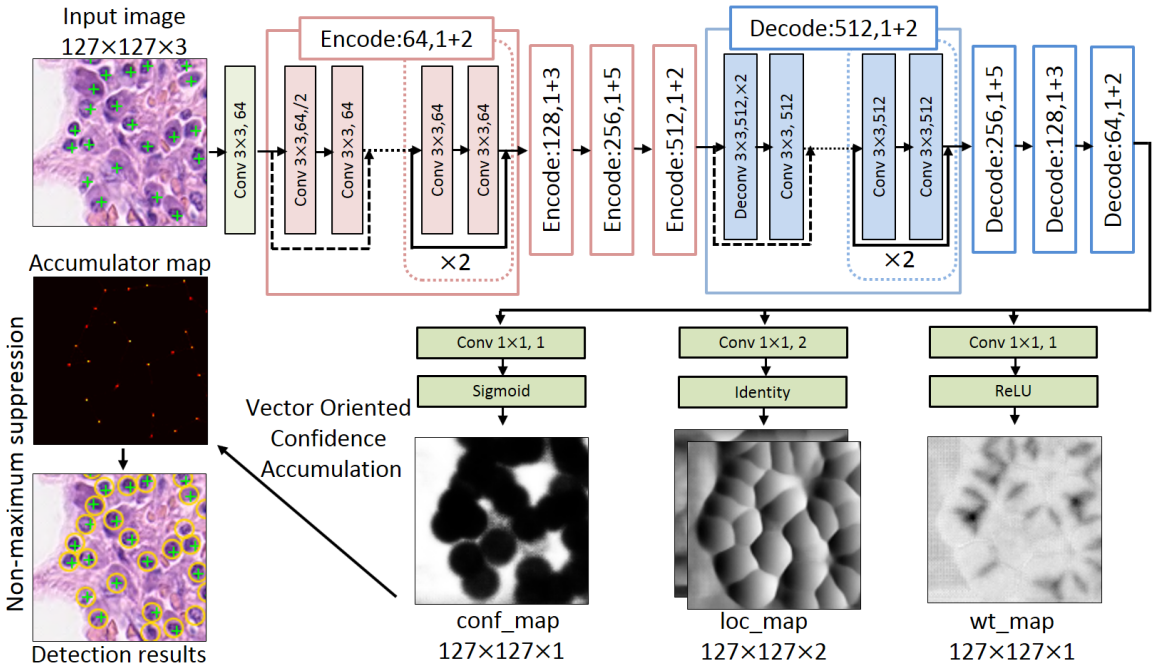

Figure 1: Residual encoder-decoder architecture of our proposed method.

We used binary cross entropy weighted by the inverse of class frequencies as the loss function for confidence score (L_{conf}). Inspired by Girshick (2015), we used smooth $l1$ loss for localization vector and weight of contribution (L_{loc}, L_{wt}) to avoid gradient explosion. The joint loss function is a linear combination of the three losses:

$$L = L_{conf} + \lambda_1 L_{loc} + \lambda_2 L_{wt} \tag{4}$$

where λ_1 and λ_2 are parameters weighting the contribution of different tasks. We kept both λ_1 and λ_2 at 1 in all of our experiments unless discussed (cf. Section 5.1).

3.2. Network architecture

Instead of computing a small patch around each pixel in the sliding-window manner, we used an FCN-like structure (Long et al., 2015) with rich features in the decoding part (Chen and Srinivas, 2016) to learn the task maps. This design shared convolutional layers and largely reduced the effective input size from the sliding-window approaches. The network abstracts and decodes distinct features for different tasks. The bottom panel of Figure 1 shows the 3 task maps. The confidence score map describes the proximity of nuclei as surrounding disks. The localization vector map is composed of two gradient images zeroed at nuclei position in both x and y dimensions. The last map correctly up-weighted the pixels at nuclei intersections. All colors were inverted for improved visualization.

Our proposed model takes input of size $127 \times 127 \times 3$ and feeds it forward to 4 encoding and 4 decoding blocks followed by 3 1×1 conv layers to produce the task maps. We used residual layers for

each block of the encoder-decoder (cf. Figure 1). Rather than max pooling, down/up-sampling was conducted within every block by 3×3 conv/deconv layers at stride 2 to retain location information. Changing the receptive field size of the last encoding block by either decreasing or increasing the number of encoding blocks degraded the detection performance in our experiments. We surmise that having a receptive field that is approximately the size of cell nuclei (16×16) on cancer slides at 20× magnification allows the network to learn higher level semantics useful for the tasks. On top of the last 1×1 conv layers, we used sigmoid activation for confidence score maps, since it is stable to our binary cross entropy loss L_{conf}. Identity function was employed as the activation to account for both negative and positive values of the regression target. For the weight of contribution map we selected ReLU as the activation to learn the positive cell counts.

3.3. Vector oriented confidence accumulation

In detection stage, the predicted task maps are combined intuitively to generate an accumulator map (cf. Figure 1). Let P be a map initialized with zeros. For every coordinate (i, j), the localization vector accumulates the weighted confidence score of pixel to the target position :

$$P[i', j'] = P[i', j'] + \hat{Wt}[i, j] \times \hat{Conf}[i, j], \text{ where } (i', j') = (i, j) + \hat{Loc}[i, j] \tag{5}$$

The confidence accumulation amplifies the stratification between fore-ground and back-ground and produces sparse response, which enhances the speed and robustness of the follow-up non-maximum suppression on P to output the final detection results.

4. Dataset and implementation details

We validated our method on the publicly available colorectal cancer dataset released by Sirinukunwattana et al. (2016)[1]. The dataset contains 100 images of size 500×500 at 20× magnification, which were cropped from 10 whole-slide images of 9 patients with colorectal adenocarcinomas. On these images there are in total 29,747 cell nuclei marked at/around the center. We randomly split the dataset for 2-fold cross validation. The image ids for each subsample is attached in Appendix A.

The network was implemented with PyTorch (Paszke et al., 2017). Images of size 127×127 were further cropped from the dataset by a uniform grid of stride 17 for translational augmentation and to match the model input size. We used batch size 8 and learning rate 0.0005 with a decay factor of 0.1 after every 3 epochs. A momentum of 0.9 was used. Input images were normalized by the mean and standard deviation calculated on the training dataset. For further data augmentation, each image has 50% chance to be flipped horizontally and then 50% chance to be flipped vertically, finally equal chances to be rotated by 0^o, 90^o, 180^o and 270^o counterclockwise. The model was trained on a single GPU within 4 hours for 10 epochs.

Table 1: Pixel-wise classification accuracy (Acc) and localization loss (L_{loc}) of training configurations with different combinations of losses.

Configuration	Fold 1		Fold 2	
	Acc	L_{loc}	Acc	L_{loc}
Conf	0.879	-	0.882	-
Loc	-	3.969	-	4.077
Conf+Loc	0.886	3.971	0.887	4.071
Conf+Loc+Wt	**0.886**	**3.967**	**0.887**	**4.061**

5. Experiments and discussion

5.1. Pixel-wise classification accuracy and localization loss

We first evaluated the effectiveness of multi-task learning. We experimented with different values of the proximity parameter r in Equation (1) and set it to 12 for all following comparisons as it gave the best F1 score in our cross validation (cf. Section 5.3). A pixel $p_I[i,j]$ is classified correctly if $\hat{Conf}[i,j] > 0.5$ and $Conf[i,j] = 1$. The pixel-wise classification accuracy (Acc) is then defined as the average accuracy of fore-ground and back-ground pixels since we have quite imbalanced sample sizes. As we mentioned before, the localization loss (L_{loc}) was calculated as the averaged sum of smooth $l1$ losses of both x and y dimensions for all pixels. In Table 1 we presented the Acc and L_{loc} of different training configurations. Conf+Loc+Wt means that all three losses were trained concurrently. Conf means that only L_{conf} was used for training. The rest configurations are defined in a similar fashion.

The results imply that the three related tasks are mutually beneficial. Especially the classification accuracy was improved if trained together with localization loss. This improvement (from 0.879 to 0.886 for Fold 1, and from 0.882 to 0.887 for Fold 2) was comparable to other optimization of the pipeline. L_{conf} and L_{wt} converges about 3 times faster than L_{loc} during training. We surmise that regression of localization vector is a more challenging objective therefore contributed more to the learning of common features. We tried various values of λ_1 in Equation 4 (while keeping λ_2 as 1): 0.1, 1, and 10, but 1 resulted in the best performance. A natural extension of our work would be experimentation with more combinations of the weighting parameters λ_1 and λ_2. It is notable that the L_{loc} almost falls under 4, which is in $l1$ form since > 1. It means that the average localization error on each dimension is only 2 pixels. This observation is consistent with the crisp accumulator maps in Figure 2 and the high localization accuracy shown in Table 2.

5.2. Accumulator map and qualitative results

We present in Figure 2 the accumulator maps and qualitative detection results generated by VOCA. For comparison, we also implemented a pixel-wise peak regression model (PR) similar to Xie et al. (2015b). The PR model replaces the multi-task maps of VOCA by a single regression map, in

1. The dataset is available at https://www2.warwick.ac.uk/fac/sci/dcs/research/tia/data

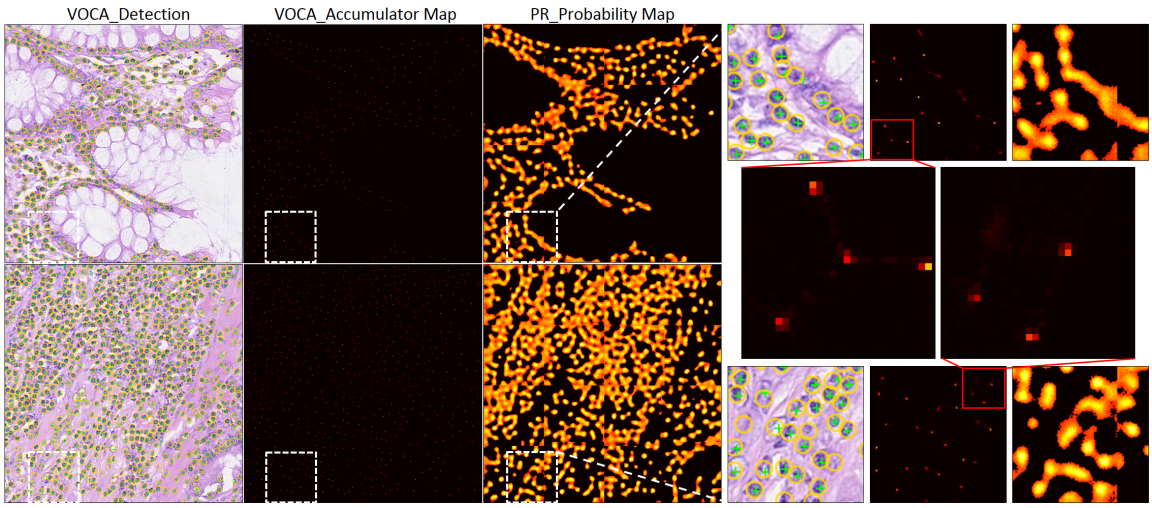

Figure 2: Accumulator maps and cell detection results of VOCA compared to peak regression (PR). The figure is best viewed on screen with magnification 400%

which the pixels are scored as $P_I(i,j) = \begin{cases} \frac{1}{(1+0.8\times||(u-i,v-j)||_2}, & \text{if } ||(u-i,v-j)||_2 < 6 \\ 0, & \text{otherwise} \end{cases}$. It is a representative of several other existing methods (Chen and Srinivas, 2016; Sirinukunwattana et al., 2016; Raza et al., 2018) which also embed recognization and localization to a single map. In detection results (cf. Figure 2 left panel), the yellow circles represent the predicted location and the green crosses are ground truth annotation. Only predictions above the confidence threshold that gives the best F1 score were shown.

As shown in the zoomed-in panels in Figure 2, the predicted confidence scores (cf. *conf_map* in Figure 1) were accumulated precisely to the target locations. Pixels with high accumulated confidence are within distance of 1 to 2 pixels to the peaks, while the majority of the background becomes zero-valued after confidence "movement". Post-processing on the clean accumulator maps of VOCA is fast. For example, it speeds up non-maximum suppression whose running time is $O(ln(n))$, where n is the number of positively valued pixels. In our experiments it took on average 0.2 seconds to process each map of size 500×500, which is about 30 times as fast as on the probability maps produced by PR (cf. Figure 2 mid panel). Besides precision, nuclei localization of VOCA also showed high accuracy as most of the yellow circles (predictions) are rigorously centered at the green crosses (ground truth). The quantitative measurement of the localization accuracy will be presented in Section 5.3.

5.3. Quantitative performance and localization accuracy

Non-maximum suppression on the crisp accumulator maps produced by VOCA is not only fast but also robust. A distance threshold of 4 pixels can already suppress most of the non-peak positions. The accumulated scores within 2 pixels of a nucleus coordinate given by non-maximum suppression

Table 2: Comparison of precision, recall, F1 score and localization accuracy

Methods	Precision	Recall	F1 score	Median Distance (Q1, Q3)
LIPSyM	0.725	0.517	0.604	2.236 (1.414, 7.211)
SSAE	0.617	0.644	0.630	4.123 (2.236, 10)
SC-CNN	0.781	0.823	0.802	2.236 (1.414, 5)
SP-CNN	0.803	0.843	0.823	-
MDN	0.788	**0.882**	0.832	-
SFCN-OPI	0.819	0.874	0.834	-
RBF-CF	0.83	0.86	0.84	-
VOCA-NW	0.814	0.854	0.834	2.0 (1.414, 2.236)
VOCA	**0.831**	0.863	**0.847**	**2.0 (1.414, 2.236)**

were summed as its final score. All scores were normalized to $[0, 1]$ for each image. The predicted coordinates were then assigned to ground truth cell nuclei by Hungarian algorithm (Kuhn, 1955) according to euclidean distance to ensure that at most 1 prediction will be considered true positive for each ground truth. The predictions are regarded as true positive if and only if they are within 6 pixels of their assigned nuclei as suggested by Sirinukunwattana et al. (2016). We plotted precision-recall curves by thresholding the final scores and obtained the optimal F1 score for comparison with the existing methods validated on the same dataset (cf. Table 2). The corresponding precision and recall were also reported.

The first panel of methods (LIPSyM (Kuse et al., 2011), SSAE (Xu et al., 2016), SC-CNN (Sirinukunwattana et al., 2016)) were (re-)validated by Sirinukunwattana et al. (2016) when they published the dataset. The second panel includes the reported results on the same dataset of more recent methods described in Section 2 (SP-CNN (Tofighi et al., 2018), MDN (Koohababni et al., 2018), SFCN-OPI (Zhou et al., 2017), RBF-CF (Ahmad et al., 2018)). VOCA-non-weighted (VOCA-NW) represents our configuration Conf+Loc (cf. Table 1) in which Wt was not trained and the confidence was thus not weighted for accumulation. "-" means the score is not available from the original paper.

VOCA achieved the best detection performance with F1 score as 0.847. It tends to have higher precision than the other methods at similar recall, which we surmise is caused by its amplification of the stratification between fore-ground and back-ground by confidence accumulation. As Wt didn't help the training (cf. Table 1), the improved performance of VOCA over VOCA-NW should come from the compensatory upweighting for pixel sharing during confidence accumulation. Theoretically VOCA-NW gives lower confidence scores for packed cells, since only a portion of the pixels at their intersections (the dark areas in the Wt map in Figure 2) are accumulated to them (illustrated in Appendix B). At certain threshold these cells will be filtered out as background by VOCA-NW while they can be correctly detected by VOCA.

We measured the same metrics as Sirinukunwattana et al. (2016) to quantitatively describe the accuracy of nuclei localization of VOCA. The Euclidean distance between each pair of ground truth and its assigned prediction was recorded for both folds of cross validation. The median, 1st quartile

and 3rd quartile of the distribution of the distances were reported. We emphasize again that the accurate coordinates of cell nuclei are the prerequisite for many downstream applications, such as tumor micro-environment analysis, and that low accuracy cell localization would result in accumulated error which hinders these tasks. Considering the radius of a cell nucleus is only around 6 to 12 pixels at 20× magnification, localization error of 5 pixels like Sirinukunwattana et al. (2016) may still introduce unignorable problems. VOCA explicitly learns nuclei localization via deep features and significantly reduced the error of 75% of the predictions to below 2.236 pixels.

6. Conclusion

In this paper, we proposed a novel deep learning algorithm called vector oriented confidence accumulation (VOCA) for large scale cell detection on histopathology images. The algorithm concurrently learns pixel-wise classification, localization and weight of contribution tasks that combine into an accumulator map which describes profoundly accurate and precise nuclei locations. Extensive experiments on a public cell detection dataset of colon cancer validated the efficacy of our proposed frame work and proved high detection performance and exceptional localization accuracy compared to the state-of-the-art, which implies high potential of a robust decision support application for various clinical and research purposes.

Acknowlegements

This work was supported by the Warren Alpert Foundation Center for Digital and Computational Pathology at Memorial Sloan Kettering Cancer Center, the NIH/NCI Cancer Center Support Grant P30 CA008748, Weill Cornell Graduate School of Medical Sciences and the Tri-I Computational Biology and Medicine Program.
T.J.F. is the chief scientific officer, co-founder, and equity holder of Paige.AI.

References

Asif Ahmad, Amina Asif, Nasir Rajpoot, Muhammad Arif, et al. Correlation filters for detection of cellular nuclei in histopathology images. *Journal of medical systems*, 42(1):7, 2018.

Yousef Al-Kofahi, Wiem Lassoued, William Lee, and Badrinath Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Transactions on Biomedical Engineering*, 57(4):841–852, 2010.

Sahirzeeshan Ali and Anant Madabhushi. An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery. *IEEE transactions on medical imaging*, 31(7):1448–1460, 2012.

Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman. Learning to detect cells using non-overlapping extremal regions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 348–356. Springer, 2012.

Jianxu Chen and Chukka Srinivas. Automatic lymphocyte detection in h&e images with deep neural networks. *arXiv preprint arXiv:1612.03217*, 2016.

Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 411–418. Springer, 2013.

Eric Cosatto, Matt Miller, Hans Peter Graf, and John S Meyer. Grading nuclear pleomorphism on histological micrographs. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–410. Springer, 2013.

Thomas J. Fuchs, Johannes Haybaeck, Peter J. Wild, Mathias Heikenwalder, Holger Moch, Adriano Aguzzi, and Joachim M. Buhmann. Randomized Tree Ensembles for Object Detection in Computational Pathology. In *Advances in Visual Computing: Part I*, ISVC '09, pages 367–378, Las Vegas, Nevada, 2009. ISBN 978-3-642-10330-8. doi: http://dx.doi.org/10.1007/978-3-642-10331-5_35. URL http://dx.doi.org/10.1007/978-3-642-10331-5_35.

Ross Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

Jisoo Jeong, Hyojin Park, and Nojun Kwak. Enhancement of ssd by concatenating feature maps for object detection. *arXiv preprint arXiv:1705.09587*, 2017.

Navid Alemi Koohababni, Mostafa Jahanifar, Ali Gooya, and Nasir Rajpoot. Nuclei detection using mixture density networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 241–248. Springer, 2018.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. ISSN 1931-9193. doi: 10.1002/nav.3800020109. URL http://dx.doi.org/10.1002/nav.3800020109.

Manohar Kuse, Yi-Fang Wang, Vinay Kalasannavar, Michael Khan, and Nasir Rajpoot. Local isotropic phase symmetry measure for detection of beta cells and lymphocytes. *Journal of pathology informatics*, 2, 2011.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 850–855. IEEE, 2006.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Workshop Autodiff*, 2017.

Shan E Ahmed Raza, Khalid AbdulJabbar, Mariam Jamal-Hanjani, Selvaraju Veeriah, John Le Quesne, Charles Swanton, and Yinyin Yuan. Deconvolving convolution neural network for cell detection. *arXiv preprint arXiv:1806.06970*, 2018.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

K. Sirinukunwattana, S. E. A. Raza, Y. W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1196–1206, May 2016. ISSN 0278-0062. doi: 10.1109/TMI.2016.2525803.

Mohammad Tofighi, Tiantong Guo, Jairam KP Vanamala, and Vishal Monga. Deep networks with shape priors for nucleus detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 719–723. IEEE, 2018.

Mitko Veta, Paul J Van Diest, Robert Kornegoor, André Huisman, Max A Viergever, and Josien PW Pluim. Automatic nuclei segmentation in h&e stained breast cancer histopathology images. *PloS one*, 8(7):e70221, 2013.

Jelte Peter Vink, MB Van Leeuwen, CHM Van Deurzen, and G De Haan. Efficient nucleus detector in histopathology images. *Journal of microscopy*, 249(2):124–135, 2013.

Haibo Wang, Angel Cruz Roa, Ajay N Basavanhally, Hannah L Gilmore, Natalie Shih, Mike Feldman, John Tomaszewski, Fabio Gonzalez, and Anant Madabhushi. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, 1(3):034003, 2014.

W. Xie, J. A. Noble, and A. Zisserman. Microscopy cell counting with fully convolutional regression networks. In *MICCAI 1st Workshop on Deep Learning in Medical Image Analysis*, 2015a.

Yuanpu Xie, Fuyong Xing, Xiangfei Kong, Hai Su, and Lin Yang. Beyond classification: structured regression for robust cell detection using convolutional neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 358–365. Springer, 2015b.

Jun Xu, Lei Xiang, Qingshan Liu, Hannah Gilmore, Jianzhong Wu, Jinghai Tang, and Anant Madabhushi. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE transactions on medical imaging*, 35(1):119–130, 2016.

Yanning Zhou, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Sfcn-opi: Detection and fine-grained classification of nuclei using sibling fcn with objectness prior interaction. *arXiv preprint arXiv:1712.08297*, 2017.

Appendix A. Image ids for each subsample

Subsample 1: 6, 8, 10, 11, 13, 17, 18, 19, 20, 21, 23, 25, 26, 27, 28, 29, 32, 33, 39, 41, 42, 45, 46, 47, 48, 49, 51, 53, 55, 56, 59, 60, 63, 65, 67, 69, 70, 75, 76, 78, 79, 84, 86, 87, 92, 93, 95, 96, 98, 100

Subsample 2: 1, 2, 3, 4, 5, 7, 9, 12, 14, 15, 16, 22, 24, 30, 31, 34, 35, 36, 37, 38, 40, 43, 44, 50, 52, 54, 57, 58, 61, 62, 64, 66, 68, 71, 72, 73, 74, 77, 80, 81, 82, 83, 85, 88, 89, 90, 91, 94, 97, 99

Appendix B. Pixel sharing during confidence accumulation