Cell Counting and Segmentation of Immunohistochemical Images in the Spinal Cord: Comparing Deep Learning and Traditional Approaches

Bau Pham, Bilwaj Gaonkar, William Whitehead, Steven Moran, Qing Dai, Luke Macyszyn, and V. Reggie Edgerton

***Abstract*—Estimation of cell nuclei in images stained for the c-fos protein using immunohistochemistry (IHC) is infeasible in large image sets. Use of multiple human raters to increase throughput often creates variance in the data analysis. Machine learning techniques for biomedical image analysis have been explored for cell-counting in pathology, but their performance on IHC staining, especially to label activated cells in the spinal cord is unknown. In this study, we evaluate different machine learning techniques to segment and count spinal cord neurons that have been active during stepping. We present a qualitative as well as quantitative comparison of algorithmic performance versus two human raters. Quantitative ratings are presented with cell-count statistics and Dice (DSI) scores. We also show the degree of variability between multiple human raters' segmentations and observe that there is a higher degree of variability in segmentations produced by classic machine learning techniques (SVM and Random forest) as compared to the newer deep learning techniques. The work presented here, represents the first steps towards addressing the analysis time bottleneck of large image data sets generated by c-fos IHC staining techniques, a task that would be impossible to do manually.**

I. INTRODUCTION

Identifying the location and functionality of neurons in the biological neural networks responsible for behavioral tasks provide insight into how to treat neurological diseases that affect them. An example would be enhancing the efficacy of spinal epidural stimulation to facilitate stepping and standing in spinal cord injured patients. The ability to spatially map the spinal cord's locomotor circuit would provide new targets and strategies for electrical stimulation [1]. Neuroscientists can use tissue processing techniques like immunohistochemistry (IHC) to understand the anatomy and function of biological neural networks.

IHC is a tissue processing technique done on thin sections of tissue (<50 μm) that stains cellular proteins using specialized antibodies. IHC can reveal anatomical, functional, and connectivity properties of spinal cord neurons depending on the target protein and its location in the neuron. In this study we look at the c-fos protein, a biomarker for neural activation located in the cell nucleus. The images in this paper show c-fos expression in mouse spinal cord neurons activated during quadrupedal stepping on a treadmill for 30 minutes. Typically, analysis of IHC images requires manual segmentation, which provides accurate analysis that is robust to image and IHC staining quality, but it consumes a lot of time. Multiple raters can decrease analysis time, but adds variance to the data analysis because different raters have different criteria for what constitutes a positive stain [2]. Recently, researchers have explored the use of machine learning tools to analyze IHC images, however, these tools have not been applied to cell counting in the spinal cord.

Automatic cell-counting in pathological images first started by using algorithms based on intensity thresholding, edge detection, template matching, and active shape models [3-7]. Machine learning techniques like support vector machines (SVM), random forests (RF), k-means clustering, and fuzzy c-means algorithms were explored later [6, 8-11]. The next generation of automatic segmentation algorithms utilizes deep learning techniques for a variety of biomedical image analysis such as mitosis detection, epithelial tumor nuclei identification, brain tumor classification, glioma grading, and segmentation of a variety of tissue including neurons, colon glands, nuclei, and epithelium [12]. Recently, analysis of IHC images have explored the use deep learning methods like convolutional neural networks (CNN) for segmentation [13].

Previous uses of machine learning and deep learning for IHC analysis has focused on immune cells and retinal cells, but have neither been used to segment neurons expressing c-fos, nor have they been used to analyze the spinal cord. We compare the performance of several different machine learning techniques like multi-scale fully convolutional networks (FCN), U-net, SVM, and RF's with DSI scores of .784, .621, .825, and .821 respectively. This work shows that machine learning algorithms trained by a particular rater, are highly biased to agree with that rater; even more so than a second human rater agrees with the first one. The high variability of human rater generated cell counts is a particularly challenging aspect of IHC cell nuclei segmentation.

The work presented in this paper lays the groundwork for solving the data analysis bottleneck of next-generation tissue processing techniques. This will be especially important, once newer techniques like CLARITY, which generate large data sets containing hundreds or thousands of images get mainstreamed. The extension of the algorithms presented here to large CLARITY data is reviewed in the discussion.

Bau Pham is with the Department of Bioengineering at UCLA, Los Angeles, CA, 90024 USA (phone number: 337-257-7048, e-mail: bphamjr@gmail.com)

Bilwaj Gaonkar and Luke Macyszyn are with the Department of Neurosurgery at UCLA, Los Angeles, CA, 90024 USA

William Whitehead and Steve Moran are with the Department of Electrical Engineering at UCLA, Los Angeles, CA 90024 USA

Qing Dai is with the Department of Biochemistry at UCLA, Los Angeles, CA, 90024 USA

V. Reggie Edgerton is with the Department of Integrative Biology and Physiology at UCLA, Los Angeles, CA, 90024 USA

978-1-5386-3646-6/18/$31.00 ©2018 IEEE

II. METHODS

A. Dataset

We train and test the machine learning algorithms on images from the lumbar region of the mouse spinal cord, specifically the L4 segment. The tissues were cut into 30 µm thick cross sections and stained for c-fos by immunohistochemistry (IHC). Positive c-fos stains mark activated neurons during a 30 minute session of quadrupedal stepping on a treadmill. Images taken had a 1200 x 1600 pixel resolution. The training set composed of 20 images from 4 different mice with varying degrees of brightness and image quality. These variations arise from different IHC trials, varied microscope settings, and different qualities of stepping behavior. The training set was intentionally designed this way to best capture the variable nature of behavioral and histological data sets. The testing set composed of 15 images from 3 different animals independent of the training set. A second testing set, created by a second human rater, included 6 images from the same 3 animals as the first testing set. This second testing was used to demonstrate variability between human raters and to test the generalizability of the automated techniques when compared to multiple human raters.

In our analysis, we compare four different machine learning techniques: Support vector machine (SVM), Random forest (RF), U-net, and Multi-scale fully convolutional network (FCN).

B. Classic Machine Learning Techniques

We use two classic machine learning techniques, SVM and RF. For both approaches, histogram equalization followed by a convolution with a Ricker wavelet was used to enhance the signal to noise ratio. The resulting image underwent intensity thresholding followed by a size filter to find candidate patches (40x40 pixels) for classification. This was done to decrease the number of patches to be analyzed compared to the total number of 40x40 patches that could be extracted from a 1200 x 1600 image. This patch size was chosen so that only one c-fos positive stain could fit in one patch. Shape, MR8 texture, and histogram of oriented gradients (HoG) features were extracted from the patches.

Shape features included solidity, orientation, diameter, area, eccentricity, convex area, major axis length, minor axis length, and extent. Textures features were based on the MR8 filter banks which include 36 bar and edge filters, a Gaussian filter, and a Laplacian of Gaussian filter. The eight highest responses are extracted to maintain rotation invariance. Histogram of Oriented Gradients (HoG) [14] features were also extracted. Combined, these features were input to SVM and RF for binary classification. We trained a linear SVM model using the default settings in LibSVM [15] and the RF algorithm using 200 bagged classification trees.

C. U-nets

We include a deep learning architecture known as the U-net in our analysis. U-nets have been successfully used in segmentation of biomedical images. Our U-Net implementation closely follows the original one from [16].

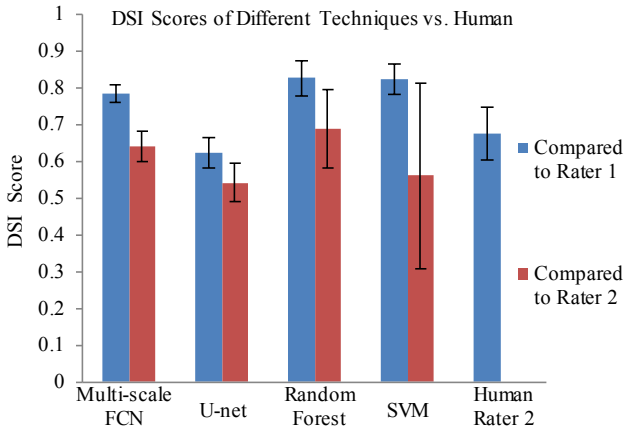

Figure 1. Performance of different segmenting techniques compared to human rater 1 (blue bar) and human rater 2 (red bar). Last blue bar on the right denotes inter-rater variability.

D. Multi-scale network

We employ the use of a multi-scale cascade of fully convolutional neural networks (FCN). Each network in the cascade operates at a different scale and its results are fed as input, in addition to the original image, to the subsequent network in the cascade. The first network, or stage, delineates anatomy on a $1/8^{th}$ scale version of the image. When the image is scaled down, large contextual features in the image shrink in dimension, allowing the neural networks to interpret large image features without having to process more pixels. The first stage's segmentation result is scaled up by a factor of two, and is added alongside the original image to become part of the input to the next stage. The second stage segments the image at 1/4th scale and the pattern continues. The second stage adds its output to the input for the third stage (1/2 scale), which then passes its output to a full-scale segmentation network that makes the final fine-scaled prediction. This method of chaining networks together is inspired by work presented by Eigen [17] for scene segmentation challenges.

Each stage is a FCN consisting of an initial pooling layer, six 3x3 convolutional layers, and a final 1x1 convolutional layer for output. Inspired by the Inception architecture [18], each of the main 3x3 convolutional layers consists of 1x1 and 3x3 convolutions in parallel. These two convolutions have an equal number of features, operate on the entire input feature space, and output half of the total output features. When results from the two convolutions are concatenated, the output dimensions are the same as the input dimensions. It is also important to understand each stage as a pixel-wise segmentation network [19]. Unlike SVM and RF, we used raw images as input to the multi-scale network and U-net with minimal to no preprocessing (e.g. histogram equalization)

III. RESULTS

Multi-scale FCN, U-net, RF, and SVM all show similar segmentation performances with Dice scores of .785, .621, .825, and .821 respectively. The results are summarized in (Fig. 1). The mean value of both Dice scores and cell counts is generally higher for the Deep Learning techniques. However, the differences between either technique is not

Machine Learning Method	Compared to Rater 1		Compared to rater 2	
	Recall	Precision	Recall	Precision
SVM	0.795	0.860	0.429	0.822
Random Forest	0.835	0.825	0.540	0.973
Multi-scale FCN	0.836	0.835	0.635	0.927
U-net	0.811	0.598	0.672	0.700

Table 1. Precision and recall for machine learning techniques when compared to human rater 1 or human rater 2 as the ground truth

statistically significant. The variance in performance when compared to human rater 2 is more in classical techniques (RF and SVM) as compared to multi-scale FCN and U-net (Fig. 1). The similarity of all the automated techniques vs. human rater 2 shows that the machine learning techniques have acquired human rater 1's biases when it comes to the criteria for what constitutes a positive c-fos stain; even though a cross validation based paradigm was employed during evaluation. Precision and recall for cell count values have been presented in Table 1.

Qualitatively, deep learning techniques are able to recognize cells in distorted c-fos; where RF and SVM can sometimes miss the identification of nuclei (Fig. 2).

IV. Discussion and Future Work

The results shown here has compared the performance of multi-scale fully convolutional networks (FCN), U-net, random forests (RF), and support vector machines (SVM) to segment images of 30 μm thick tissue slices stained for c-fos by immunohistochemistry (IHC). The variability of manual segmentation between human raters highlights the trade-off between decreasing analysis time and increasing the variability of the analyzed data when using multiple human raters. Using trained machine learning (ML) techniques for image segmentation guarantees that the criteria for what constitutes a positive signal stays the same for all images analyzed, while also decreasing the segmentation time by orders of magnitude compared to using multiple human raters. These automated analyses also do not suffer from fatigue or attention deficits that affect human raters.

The decrease in recall (Table 1) and DSI scores (Fig. 1) when going from rater 1 to rater 2 demonstrates that trained algorithms have acquired the bias of human rater 1 for which it has been trained on. This discrepancy highlights the possibility of needing to train machine learning algorithms on multiple raters to compensate for the differences in criteria for multiple raters. The increased variance in DSI scores when going from rater 1 to rater 2 of RF and SVM versus deep learning (DL) techniques shows that the deep learning techniques may be capturing a more effective underlying representation. This is further highlighted by the better precision of DL methods when going from rater 1 to rater 2. Future work will focus on refining and developing DL methods. DL methods simplify image segmentation by using the raw image. Secondly, there is an intense focus on deep learning, within the broader community, leading to the availability of several open source nature DL packages.

The use of automated segmentation greatly facilitates the analysis of c-fos IHC staining in the spinal cord. However, the automated analysis compared here can extend to other proteins stained by IHC as long as there is sufficient training data. Segmentation of c-fos alone reveals the spatial distribution of neurons in the biological neural networks responsible for particular tasks (i.e. walking) and thus reveal potential target locations for treatments of disorders that affect them (i.e. electrical stimulation for spinal cord injury). C-fos staining combined with other IHC stains can further reveal how and where the biological network as a whole delegates subsets of functionality to accomplish a behavior. For example, combining c-fos staining with staining for inhibitory or excitatory neural markers can reveal how and where the neural network uses excitation and suppression to accomplish the goal of stepping. This provides a richer understanding of biological neural networks that combines physical architecture with functionality. This deeper understanding has the potential to allow us to manipulate or artificially recreate these networks.

The current study uses automated machine learning techniques on thinly sliced tissue sections, but this has the potential to quickly analyze images generated from next-generation tissue processing techniques like CLARITY. CLARITY renders thick sections of tissue (>500 μm) optically transparent. The transparency of the tissue allows a confocal or light sheet microscope to create 3-D reconstructions by take hundreds to thousands of images through the whole thickness of the tissue at sub-micron steps. CLARITY-cleared tissue sections can also undergo IHC staining despite the increased thickness of tissue sections. Though this processing technique provides a breakthrough in imaging biological neural networks, there is no way to analyze the large amount of data generated in a timely manner, which limits the potential that this processing technique can offer. Automated DL techniques can analyze these image sets in a timely manner, and will aid in future work on tract tracing to identify the connectivity of biological neural networks that would be impossible to analyze manually. This would allow us to accomplish the goals outlined in the previous paragraph, except on a more global scale.

Acknowledgment

The authors would like to acknowledge Dr. Niranjala Tillakaratne for her contribution as human rater 2. Without her effort we would not be able to gauge the variability between multiple human raters.

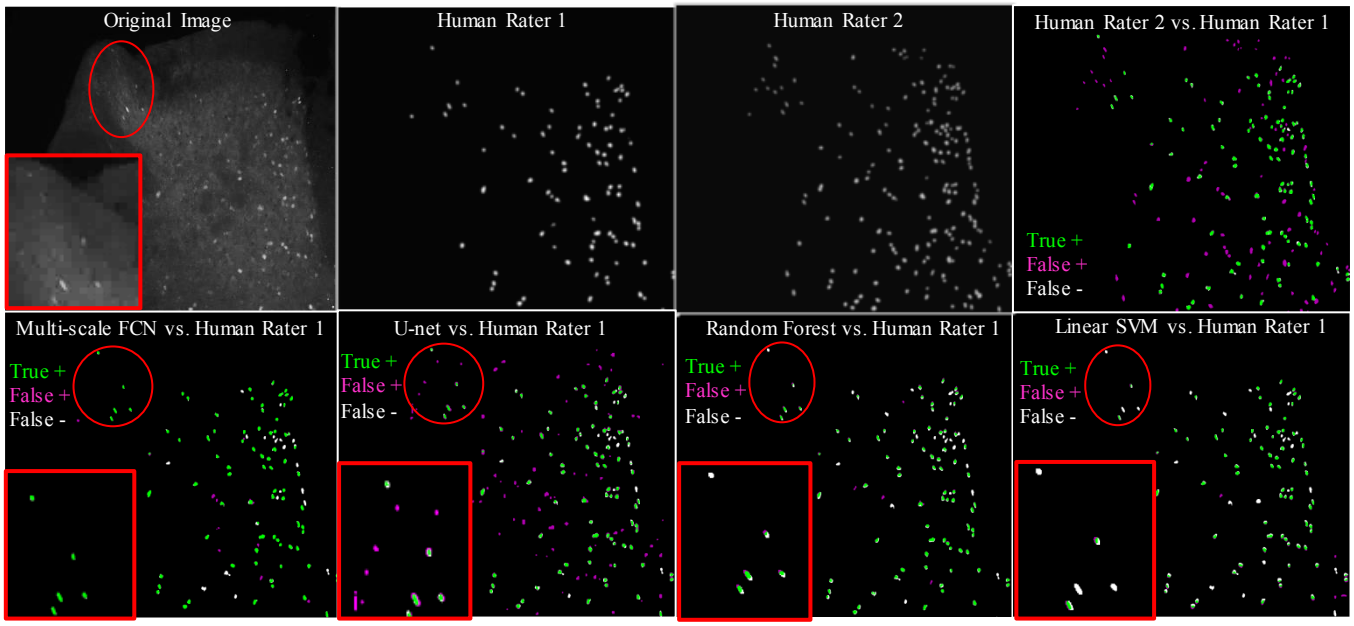

Figure 2. Qualitative results of machine learning techniques and human raters. Notice the variability between the segmenation of human rater 1 and 2. Red circled represent an area where c-fos has been distorted due to a mounting error and is enlarged in the red box. These red boxes highlight that deep learning techniques are able to detect this distortion without prior training while the classic machine learning techniques only properly identify a few of them.

REFERENCES

[1] S. Harkema *et al.*, "Effect of epidural stimulation of the lumbosacral spinal cord on voluntary movement, standing, and assisted stepping after motor complete paraplegia: a case study," *Lancet,* vol. 377, no. 9781, pp. 1938-47, Jun 4 2011.

[2] M. Lacroix-Triki *et al.*, "High inter-observer agreement in immunohistochemical evaluation of HER-2/neu expression in breast cancer: a multicentre GEFPICS study," (in eng), *Eur J Cancer,* vol. 42, no. 17, pp. 2946-53, Nov 2006.

[3] T. Yang, W. Peng, X. Li, and Y. Wang, "A Robust IHC Color Image Automatic Segmentation Algorithm," Berlin, Heidelberg, 2014, pp. 319-328: Springer Berlin Heidelberg.

[4] B. v. Ginneken, A. F. Frangi, J. J. Staal, B. M. t. H. Romeny, and M. A. Viergever, "Active shape model segmentation with optimal features," *IEEE Transactions on Medical Imaging,* vol. 21, no. 8, pp. 924-933, 2002.

[5] J. Dong, J. Li, A. Fu, and H. Lv, "Automatic Segmentation for Ovarian Cancer Immunohistochemical Image Based on YUV Color Space," in *2010 International Conference on Biomedical Engineering and Computer Science*, 2010, pp. 1-4.

[6] S. Di Cataldo, E. Ficarra, A. Acquaviva, and E. Macii, "Automated segmentation of tissue images for computerized IHC analysis," *Computer Methods and Programs in Biomedicine,* vol. 100, no. 1, pp. 1-15, 2010/10/01/ 2010.

[7] C. Chen, W. Wang, J. A. Ozolek, and G. K. Rohde, "A flexible and robust approach for segmenting cell nuclei from 2D microscopy images using supervised learning and template matching," (in eng), *Cytometry A,* vol. 83, no. 5, pp. 495-507, May 2013.

[8] P. Shi, J. Zhong, J. Hong, R. Huang, K. Wang, and Y. Chen, "Automated Ki-67 Quantification of Immunohistochemical Staining Image of Human Nasopharyngeal Carcinoma Xenografts," *Sci Rep,* vol. 6, p. 32127, Aug 26 2016.

[9] J. Oscanoa, F. Doimi, R. Dyer, J. Araujo, J. Pinto, and B. Castaneda, "Automated segmentation and classification of cell nuclei in immunohistochemical breast cancer images with estrogen receptor marker," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 2399-2402.

[10] F. Mualla, S. Scholl, B. Sommerfeldt, A. Maier, and J. Hornegger, "Automatic Cell Detection in Bright-Field Microscope Images Using SIFT, Random Forests, and Hierarchical Clustering," (in eng), *IEEE Trans Med Imaging,* vol. 32, no. 12, pp. 2274-86, Dec 2013.

[11] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman, "Learning to Detect Cells Using Non-overlapping Extremal Regions," Berlin, Heidelberg, 2012, pp. 348-356: Springer Berlin Heidelberg.

[12] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of Pathology Informatics,* Original Article vol. 7, no. 1, pp. 29-29, January 1, 2016 2016.

[13] W. Xie, J. A. Noble, and A. Zisserman, "Microscopy cell counting and detection with fully convolutional regression networks," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization,* pp. 1-10, 2016.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 886-893 vol. 1.

[15] C.-C. a. L. Chang, C.-J., " LIBSVM: A library for support vector machines. ," *ACM Trans. Intell. Syst. Technol.,* vol. 2, 3,, no. Article 27 p. 27 pages, (April 2011) 2011.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Cham, 2015, pp. 234-241: Springer International Publishing.

[17] D. Eigen and R. Fergus, "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2650-2658.

[18] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1-9.

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