Final Project Proposal: Nuclei Detection

Yining Cao
yinicao@iu.edu

Chuck Jia
jiac@iu.edu

Ruiyu Zhu
zhu52@iu.edu

March 3, 2018

1 Objective and Significance

The objective of our project is to create an algorithm that automates the detection of nuclei from medical images. Nuclei are important subjects in medical research and diagnoses, as they contain most of a cell's genetic material, organized as DNA molecules. Identifying cell nuclei is the first step for most medical analyses, because it enables researchers to isolate individual cells from samples containing multiple cells. By performing further analysis on individual cell nuclei and observing cell activities and reactions to treatments, researchers can understand the underlying biological processes and use the knowledge to design new treatments or techniques. In particular, the aspect of nuclei is critical for evaluating the existence of many diseases (e.g. cancers) and their severity.

However, identifying nuclei is very time-consuming when performed by people. Automating the nuclei detection process will free researchers in medicine and biology so that they can focus on finding cures for diseases. And it may eventually speed up the cures for various diseases as well as reduce the time-to-market for new drugs, which is currently around 10 years on average.

2 Backgrounds

2.1 Basic concepts

The nucleus is a membrane-enclosed organelle found in eukaryotic cells. Most of the human body's 30 trillion cells contain a nucleus full of DNA, the genetic material that programs each cell. Nucleus' shape, size, and texture depend on the nuclei type, malignancy of the disease and the nuclei life cycle. Mitotic nuclei (MN), a stage of nuclei life circle, is another factor that may cause variations to the nucleus' appearance and increase the complexity during a nuclei detection. Among the different types of nuclei, two types are usually the object of particular

interest: lymphocyte nuclei (LN) and epithelial nuclei (EN). Epithelial cells line the outer surfaces of organs and blood vessels throughout the body, as well as the inner surfaces of cavities in many internal organs. Normal EN have nearly uniform chromatin distribution with smooth boundary. In contrast, EN in high-grade cancer cells are larger in size, may have heterogeneous chromatin distribution, irregular boundaries, referred to as nuclear pleomorphism, and clearly visible nucleoli as compared to normal EN. Lymphocyte is one of the subtypes of white blood cell in a vertebrate's immune system. LN are inflammatory nuclei having regular shape and smaller size than EN. These variations of nucleus' appearance are one of the main difficulties in nuclei detection.

2.2 Related Works

The most commonly used image processing methods in nuclei detection, according to the review paper [1], include: thresholding, morphology, region growing, watershed, active contour models (ACMs) and level sets, K-means clustering, probabilistic models (e.g. Gaussian mixture models (GMMs)) and graph cut (e.g. normalized cut (Ncut)). A large number of publications on nuclei segmentation used the above methods, separately or in combination. The simplest way to detect and segment nuclei is thresholding followed by morphological operations, see e.g. [2],[3]. This methodology reports higher performance on well-defined, preferably uniform background. Huang and Lai [4] proposed watershed and ACM-based framework for nuclei segmentation in hepatocellular carcinoma biopsy images. However, this framework are sensitive to initialization and achieves poor segmentation in case of low contrast, noisy background, and damaged/irregular nuclei. [5] proposed a method based on graph cut methods. The authors compared this method with K-means clustering and Bayesian classification methods in [6]. This method reported 95.73% segmentation accuracy as compared to K-means clustering and Bayesian classification methods which reported 93.67% and 96.47% accuracy, respectively. Kofahi et al. [7] proposed another Gcuts-based method that is initialized using response of the image to Laplacian of Gaussian (LoG) filter for segmentation of breast cancer nuclei (CN). The authors reported 86% accuracy on 25 histopathological images containing 7400 nuclei. This framework often causes oversegmentation when chromatin is highly textured and the shape of nuclei is extremely elongated. But in case of highly clustered nuclei with weak borders between nuclei, undersegmentation may occur.

The above traditional nuclei segmentation frameworks have reported good segmentation performance on LN and normal EN having regular shape, homogeneous chromatin distribution, smooth boundaries, and individual existence. However, these frameworks have poor segmentation accuracy for CN especially when CN are clustered and overlapping. When there is chromatin variations. which is common in CN, the performance is also very poor. Some nuclei segmentation frameworks tackles the challenges of heterogeneity, overlapping, and clustered nuclei by using machine learning algorithms, statistical and shape models together with classical segmentation methods.

For example, Wahlby et al. [8] addressed the problem of clustered nuclei and combined seeded

watershed on gradient magnitude images with shape-based cluster separation method to improved segmentation. The seeds were found using morphological filtering. Fatakdawala et al. [9] proposed a Expectation-Maximization (EM) driven Geodesic ACM with overlap resolution for segmentation of LN in breast cancer histopathology. EM-driven ACM can reduce the original methods' sensitivity to initialization and allows the model to focus on relevant objects of interest.

In general, most model-based approaches segment nuclei using a priori information, which may introduce a bias favoring the segmentation of nuclei with certain characteristics. More recent studies have shown that deep learning methods produce promising results in nuclei detection. Xu et al. [10] used stacked sparse autoencoder to learn a high- level representation of nuclear and non-nuclear objects in an unsupervised fashion. Xie et al. [11] proposed structural regression convolution neural networks (CNNs) capable of learning a proximity map of cell nuclei and was shown by the authors to provide more accurate detection results. Another closely related deep learning work is by Xie et al. [12], which localizes nucleus centroids through a voting scheme.

Back to our project, part of our plan, as we will specify in the next section, is to use traditional image segmentation methods to provides candidates for classifications, then we may use deep learning methods and other classification algorithms (e.g SVM) to detect the nuclei. Hence, we will try to combine image segmentation methods with deep learning and machine learning. One of our goals is to see which kind of combination provides the best performance. It could be possible that each combination has its own pros and cons, in which case we will be interested in finding out what the advantages are for each of the combinations.

3 Proposed Approach

3.1 Data Description and Collection

For this project, we will use data sets from the Kaggle competition "2018 Data Science Bowl: Find the Nuclei in Divergent Images to Advance Medical Discovery", provided by the host of the competition Booz Allen Hamilton, Inc.

Our data set consists of a train set and a test set. The train set contains 670 nuclei images, in the format of PNG files. Each image contains multiple nuclei. Also included in the train set are the mask images for the location of nuclei. For each nuclei image, multiple masks are provided, each representing the location of one individual nucleus within the image. The mask images are of the same size with the nuclei image and are in binary format, using white pixels to denote the location of a nucleus and black pixels for areas outside of that nucleus. The test set contain 65 nuclei images, but does not contain any mask images.

A preliminary examination of the data sets shows that the quality of the images and masks is

decent. All the images, taken under lab conditions, show very little noise, and the masks were given in accurate binary format.

However, although the noise is minimal in our data set, the nuclei images were taken under a variety of conditions, which will pose as difficulties in our training and predicting process. For example, our images are taken under different lighting situations, and therefore they vary in luminance levels. Different color profiles are also used across the images. Some images are grayscale images while others are color images. Besides the differences on image properties and qualities, the images cover a wide range of different types of cells and the magnification of cells also vary significantly. Because of these varying conditions, we will need to extract better features from the images in order to design a highly accurate detection algorithm.

3.2 Techniques and Details

Our task is to detect individual cell nucleus from images that contain multiple instances of cell nuclei. We will design an algorithm that achieves two goals for this task. One goal is identify areas of the image that might potentially contain cell nuclei. That is, we need to identify all the "nuclei-like" objects within the image and separate them with the "outside" areas that contain no nuclei. The other goal is to correctly detect and mark the areas that contain cell nuclei.

After examining the example images from the data sets, we realized that it is relatively hard to directly identify one particular cell nucleus out of many. Some of the difficulties include the fact that we don't know the exact number of cell nuclei in a given image, and we have no information on the size or geometric shapes of the nuclei. Therefore no existing one algorithm seems to be able to directly apply to the task. Therefore, we came up with a strategy to decompose the challenge into two separate parts. Part one is to segment each of the nuclei image into smaller areas and mark the areas that might potentially be a cell nucleus. The other one is to decide whether each segment marked as a potential cell nucleus from the previous step actually contains a cell nucleus.

3.2.1 Identify Suspect Cell Nuclei

For this task, we want to extract potential cell nuclei out of an image. We expect the algorithm to produce low or zero false negative results, i.e. to extract every single potential cell nucleus in an image. Since the goal of this step is to extract all the suspects, we are not focusing on making 100% sure decisions. Instead, for the reason we will explain later, in this step, we will focus on making sure all the suspects are extracted and under our control.

To accomplish the task, we will use techniques from the computer vision area. Based on our observation, a cell nucleus always has a relatively consistent border, and its interior has significantly different pixel values from exterior areas. Therefore we will try to recognize all the suspect cell nuclei based on the geometric and color properties of their borders and

interiors. To achieve this goal, we will be using two types of computer vision techniques: image segmentation and edge detection. To perform segmentation of the cell images, we will experiment with the thresholding method, which is one of the simpler methods, and the spectral clustering methods, such as normalized cuts. We will also apply edge detection methods in this step, in order to maximize the accuracy of the result. First, to extract the potential borders of the cell nuclei, edge detection algorithms including the Canny edge detection algorithms and the ellipse Hough transform will be used in this process. Then using the borders extracted, we will use an template matching algorithm to detect all the closed area that are nucleus-like in the image, based on the properties of the interior of the closed area. For the template matching, we plan to implement and experiment with similarity analysis in the image and Fourier spaces, as well as neural network approaches. Then we will combine results from the two approaches and obtain all the suspects for the cell nuclei.

The goal of our algorithm is not decomposing the input picture into distinct segments. It will only identify all the (potentially overlapping) segments that seem to have a nucleus-like figure. The output of this algorithm is a collection of potential nuclei. Those candidates will then be fed into the next step to decide whether they are truly a cell nucleus.

Therefore, it is acceptable to have a high false-positive error rate in this step. The false positives will be re-examined by another "real" cell nuclei identification process later and the errors will then to be "corrected". On the other hand, false-negatives in this phase will directly lead to errors in our final result, since we will not rescan any of the areas marked as non-nuclei in the following steps.

3.2.2 Determine Real Cell Nuclei

After completing the previous step, we now possess a processed data set consisting of potential cells. The data points in our data set are individual image segments that contains at most 1 nucleus. The first task is to label these data points. Comparing the processed data set, which contains segments of potential nuclei, with the provided masks in the training set, we can obtain the set of "true nuclei" (TP) and "fake nuclei" (FP) extracted by the extraction algorithm we just discussed before.

Now we have reduced the task into a binary classification problem. There are tons of great algorithms/possibilities to try. However, we still have one more step to go on top of any classification algorithms. That is to normalize the internal data. We will first be centering on the suspect cells before feeding them into a classifier.

Currently, we are planning to use support vector machine (SVM) with different kernels to see if it can accomplish such classification task. We will also try deep neural network techniques for the training process.

In order to use SVM, our plan is to convert every suspect nuclei picture into a gray-scale vector. Then we can feed the data into a SVM with different kernel functions. This classification

problem seems to be non-linear. Therefore our tier-1 candidate kernel functions include graph kernels (i.e. random walk kernel) and radial basis function kernel (RBF). We will also try with polynomial kernels to observe if a linear kernel fits better in this situation.

For the deep neutral network approach, we will first enrich our training dataset. Our strategy is to rotate/flip every record to produce multiply instances in the training set. We will carry out such operation on both true nuclei and fake nuclei such that the classifier will have more instances on non-nucleus objects. Then we can use max-pooling on the training set after converting all the pictures into a gray-scale one. Finally we will try to tune up the parameters of the neutral network manually.

3.3 Evaluation

The competition provides a detailed evaluation metric (provided below)[1]. We will use the same method to evaluate the accuracy of our approach. In addition, we will perform evaluation on both the output of the extraction algorithm, a.k.a. the internal data, and the final results to demonstrate the effect of the second classification procedure.

To determine the quality of our final result, we will evaluate using k-fold validation methods on our training set. In addition, we will also use the test data set for evaluation purposes. As the test masks are not public information until the official end of the competition, the evaluation scores for the test set will be calculated by Kaggle.com, using the same evaluation methods.

3.3.1 Provided Evaluation Metric

For the evaluation of accuracy, we use the intersection over union (IoU) method. The IoU of a detection result D and the true location T, both as local images, can be calculated as

$$IoU(D,T) = \frac{D \bigcap T}{D \bigcup T}$$

The accuracy for a detection result on one nuclei image is calculated as the mean average IoU precision at different thresholds, as described below.

A threshold is used to determine if our predicted location for one cell nucleus is an acceptable result or a "hit". For example, for a threshold of 0.5, if the IoU of our detection is greater than 0.5, then our result will considered a hit, i.e. a true positive. For a detection with an IoU score less then 0.5, the detection result for that nucleus will be considered as not acceptable, i.e. a false positive.

At each threshold value t, a precision value is calculated using the number of true positives

[1]The description of the evaluation metric below follows closely with the evaluation section in the competition webpage.

(TP), false negatives (FN), and false positives (FP), by

$$\text{precision} = \frac{TP(t)}{TP(t)+FP(t)+FN(t)}$$

The TP, FN, FP numbers will be calculated by comparing the predicted result with the ground truth objects using the IoU scores with a specific threshold. The average precision of a single image is then calculated as the mean of the above precision values at each IoU threshold:

$$\frac{1}{|\text{number of thresholds}|}\sum_{t}\frac{TP(t)}{TP(t)+FP(t)+FN(t)}$$

In the end, the accuracy returned by the competition metric is the mean taken over the individual average precision of each image in the test data set.

3.4 Expected Outcome

For the extraction algorithm, the expected outcome is a set of most or even all of the "true" cell nuclei mixed with a low number of "fake" cell nuclei. The outcome will be in the format of binary mask images. At this stage, our outcome might contain a fair number of false positives. But we expect the outcome to contain zero or low number of false negatives.

The second phase of our algorithm would filter out the false positives in the outcome of the last step and provide accurate locations for "true" cell nuclei. The outcome will also be given in the format of binary masks.

We plan to work on both the SVM approach and the deep neural network approach for our data training. In the end, we will either combine the two approaches or select the better one of the two. Therefore, they both serve as our "fall back" options should one of them fails.

4 Individual Tasks

- **Yining Cao:** Use image segmentation methods to identify suspect cell nuclei. The methods I am going to try are: thresholding, normalized cuts, watershed and k-means. They may be applied separately or in combination, based on the outcomes.
- **Chuck Jia:** Perform edge detection and template matching on nuclei images to detect local image segments that contain a single cell nucleus. The methods developed will be applied to both phases in our plan. For the first phase of searching for suspect nuclei, algorithms and techniques from computer vision will be used, including Canny edge detection algorithms and sliding windows search. For the second phase of identifying nuclei segments, I will perform template matching in image and Fourier spaces, and train models using deep neural network techniques.

- **Ruiyu Zhu:** Identify true nuclei out of the suspects. The plan is first carrying out the classification with SVM. There are a number of candidate kernels to play with. Hopefully some of them will produce a decent accuracy. My backup plan is using neural network handling the classfication problem.

References

[1] H. Irshad, A. Veillard, L. Roux and D. Racoceanu, *Methods for Nuclei Detection, Segmentation, and Classification in Digital Histopathology: A Review—Current Status and Future Potential*, IEEE Reviews in Biomedical Engineering, vol. 7, pp. 97-114, 2014.

[2] O. S. Al-Kadi, *Texture measures combination for improved meningioma classification of histopathological images*, Pattern Recog. vol. 43 no. 6 pp. 2043-2053 2010.

[3] H. Irshad, *Automated mitosis detection in histopathology using morphological and multichannel statistics features*, J. Pathol. Inform. vol. 4 pp. 10–15 May 2013.

[4] P. W. Huang Y. H. Lai, *Effective segmentation and classification for HCC biopsy images*, Pattern Recog. vol. 43 no. 4 pp. 1550-1563 2010.

[5] V.-T. Ta O. Lézoray A. Elmoataz S. Schüpp, *Graph-based tools for microscopic cellular image segmentation*, Pattern Recog. vol. 42 no. 6 pp. 1113-1125 2009.

[6] O. Lezoray H. Cardot, *Cooperation of color pixel classification schemes and color watershed: A study for microscopic images*, IEEE Trans. Image Process.¡/em¿ vol. 11 no. 7 pp. 783-789 Jul. 2002.

[7] Y. Al-Kofahi W. Lassoued W. Lee B. Roysam, *Improved automatic detection and segmentation of cell nuclei in histopathology images*, IEEE Trans. Biomed. Eng. vol. 57 no. 4 pp. 841-852 Apr. 2010.

[8] C. Wahlby I. M. Sintorn F. Erlandsson G. Borgefors E. Bengtsson, *Combining intensity edge and shape information for 2D and 3D segmentation of cell nuclei in tissue sections*, J. Microsc. vol. 215 no. 1 pp. 67-76 2004.

[9] H. Fatakdawala J. Xu A. Basavanhally G. Bhanot S. Ganesan M. Feldman J. E. Tomaszewski A. Madabhushi, *Expectation maximization-driven geodesic active contour with overlap resolution (EMaGACOR): Application to lymphocyte segmentation on breast cancer histopathology*, IEEE Trans. Biomed. Eng. vol. 57 no. 7 pp. 1676-1689 Jul. 2010.

[10] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madab- hushi, *Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images*, Medical Imaging, IEEE Transactions on, vol. PP, no. 99, pp. 1–1, 2015.

[11] Y. Xie, F. Xing, X. Kong, H. Su, and L. Yang, *Beyond classifica- tion: Structured regression for robust cell detection using convolutional neural network*, in Medical Image Computing and Computer-Assisted InterventionMICCAI 2015. Springer, 2015, pp. 358–365.

[12] Y. Xie, X. Kong, F. Xing, F. Liu, H. Su, and L. Yang, *Deep voting: A robust approach toward nucleus localization in microscopy images*, in Medical Image Computing and Computer-Assisted Intervention– MICCAI 2015. Springer, 2015, pp. 374–382.