33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Supervised Learning Framework For Screening Nuclei in Tissue Sections

Kaustav Nandy, Prabhakar R. Gudla, Ryan Amundsen, Karen J. Meaburn, Tom Misteli and Stephen J. Lockett

Abstract— **Accurate segmentation of cell nuclei in microscope images of tissue sections is a key step in a number of biological and clinical applications. Often such applications require analysis of large image datasets for which manual segmentation becomes subjective and time consuming. Hence automation of the segmentation steps using fast, robust and accurate image analysis and pattern classification techniques is necessary for high throughput processing of such datasets. We describe a supervised learning framework, based on artificial neural networks (ANNs), to identify well-segmented nuclei in tissue sections from a multistage watershed segmentation algorithm. The successful automation was demonstrated by screening over 1400 well segmented nuclei from 9 datasets of human breast tissue section images and comparing the results to a previously used stacked classifier based analysis framework.**

I. Introduction

Many high throughput biological and clinical applications require selection of objects of interest in large microscope image datasets that have been segmented with a high degree of accuracy and confidence. Manual segmentation of such large datasets is both subjective and time consuming, making it essential to automate the processing. One such application is spatial analysis of gene localization in interphase nuclei using fluorescence in situ hybridization (FISH) technique [1], [2]. In these studies, it has been shown that localization of certain genes in interphase nuclei has implications for their function. Moreover, gene localization of certain genes is different in normal and cancerous tissues, suggesting a diagnostic value for gene localization. In this application, accurate segmentation of cell nuclei is a prerequisite for drawing significant biological and diagnostic conclusions. The task of segmenting nuclei for this application is uniquely different to other tasks. On the one hand, many more nuclei are imaged than are needed for such analysis, enabling us to emphasize highly accurate segmentation of a subset of nuclei rather than attempting to segment as many nuclei as possible. On the other hand, there is considerable variation in size, morphological and textural features of the nuclei because of the inherent variations between tissue samples and truncation of the nuclei by the physical sectioning of the tissue. High texture in the nuclear regions makes it difficult to distinguish between the boundary and internal intensity variations. Variation in morphological cues used by image analysis and pattern recognition algorithms [3] to segment nuclei complicates identification of well segmented nuclei. A fast, robust and accurate automatic processing pipeline was presented at this conference in the year 2009 [4] which identifies a subset of the objects of interest (cell nuclei) from the microscope images with a high degree of confidence. Here, we present a series of advancements to this pipeline in terms of improved segmentation, measuring more features of segmented objects and replacing the stacked classifier with an artificial neural network (ANN).

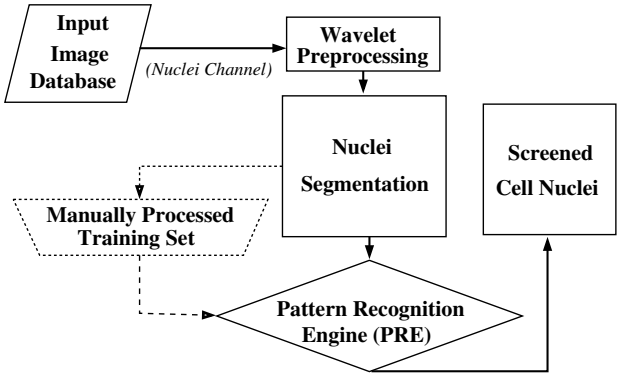


Fig. 1. Processing pipeline

Nuclei segmentation [5], [6], [7], [8] in tissue images is the first step in the workflow. The segmentation part of the pipeline incorporates multiscale edge enhancement and multistage watershed algorithms. However due to the aforementioned difficulties, 100% segmentation accuracy is not achievable. Consequently an ANN was trained on a subset of the data (25%) and then used automatically to identify with a certain degree of confidence a subset of the segmented objects.

Although the basic building blocks of the pipeline were used out of the box, the novelty of the method lies in the fact that the basic blocks have been put together in an unique and innovative way to solve an extremely challenging

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), NCI, Center for Cancer Research and by a Department of Defense Breast Cancer Idea Award to Tom Misteli.

Kaustav Nandy, Prabhakar R. Gudla and Stephen J. Lockett are with Optical Microscopy and Analysis Laboratory, Advanced Technology program, SAIC-Frederick, Inc., NCI-Frederick, Frederick, MD 21702 nandyk@mail.nih.gov, gudlap@mail.nih.gov, locketts@mail.nih.gov

Ryan Amundsen is with Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor ryan.j.amundsen@gmail.com

Karen J. Meaburn and Tom Misteli are with Cell Biology of Gene Expression, National Cancer Institute, National Institutes of Health, Bethesda, MD USA 20892 meaburnk@mail.nih.gov, mistelit@mail.nih.gov

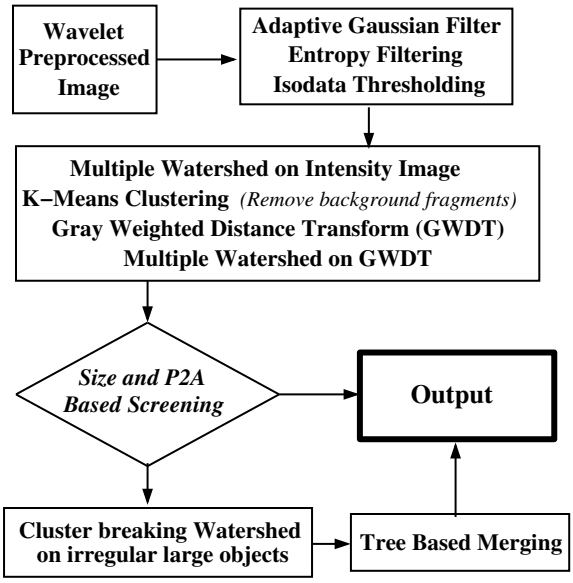


Fig. 2. Multistage Watershed Segmentation

segmentation and screening task.

II. Image Data, Nuclei Segmentation and Classification Pipeline

Sample preparation and image acquisition are as described in reference [4]. We reduced 178 3D images from 9 datasets (D1 - D9) consisting of normal and cancerous breast tissue sections to 2D images using maximum intensity projection.

Fig. 1 shows the block diagram of the segmentation and classification framework. The preprocessing step enhanced the contrast of the nuclei boundaries and the enhanced images were the input to the segmentation algorithm (Fig. 2). Output of the segmentation algorithm was classified using a supervised pattern recognition engine to identify well segmented nuclei having reasonable boundary accuracy.

A. Wavelet Preprocessing

The preprocessing step used wavelet based enhancement of the object boundaries using LastWave toolbox [9]. It involved storing the *edges* in the image using a chain coded *extrema* representation followed by selectively enhancing edges in different spatial scales using an user-defined factor. Though this step accentuated the inside texture of the nuclei (compare Figs. 3(a) & 3(b)), the advantage offered by the boundary enhancement overshadowed this shortcoming.

B. Multistage Watershed Segmentation

Fig. 2 shows the mutistage watershed based segmentation algorithm which replaced previously used hybrid levelset-watershed algorithm [4] for nuclei segmentation. Wavelet preprocessed images were first filtered using an edge preserving adaptive Gaussian filter to reduce noise and texture variations. Next, entropy based filtering followed by isodata thresholding was used to identify the foreground region. Morphological operations and size based screening was used to remove small objects resulting from noisy background

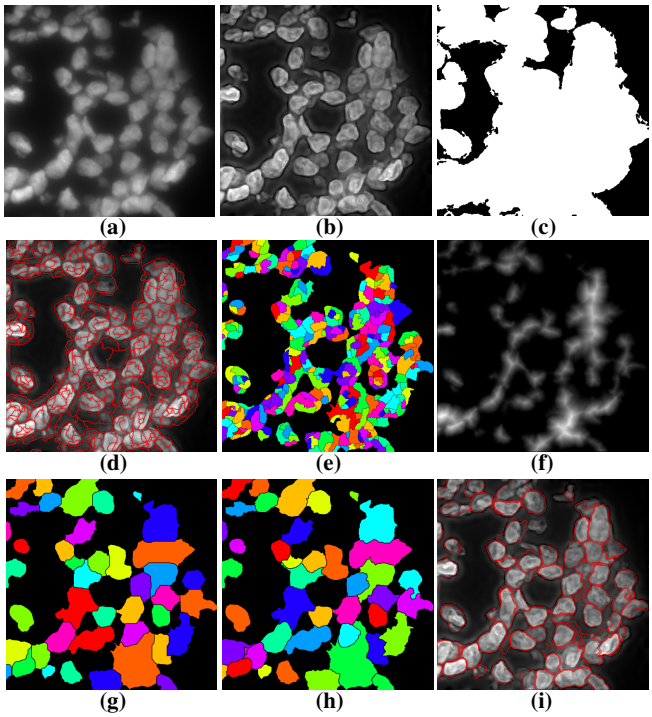


Fig. 3. (a) Original DAPI channel maximum intensity projection (MIP). (b) Wavelet enhanced DAPI channel. (c) Entropy filtered and isodata thresholded image after binary operations and size filtering. (d) Seeded watershed output. (e) Remaining watershed fragments after rejecting background fragments. (f) Gray weighted distance transform output. (g) Watershed output on the gray weighted distance transformed image. (h) Merged output of (e) and (g). (i) Final segmentation output after the cluster breaking watershed and tree based merging.

and texture within the nuclei (Fig. 3(c)). To overcome the problems of intensity variation and multiple maxima identification, due to texture, multiple runs of an intensity based seeded watershed algorithm [10] were used to obtain an initial segmentation of the nuclei (Fig. 3(d)). Seeds to initiate the watershed segmentation were identified using an extended-maxima transform [11]. Predominant watershed lines were retained as prominent edges (Fig. 3(d)). To remove the background fragments, we performed k-means intensity based clustering of the watershed fragments and rejected fragments in the lowest intensity cluster (Fig. 3(e)).

To merge the fragments from the watershed we took advantage of the expected morphological structure of nuclei by using a gray weighted distance transform (GWDT) (Fig. 3(f)). The GWDT helped identify and seed high intensity nuclei regions, while taking advantage of the structure of the foreground area. Multiple runs of the seeded watershed were used to segment the distance image (Fig. 3(g)). Intensity based watershed fragments were assigned the label of the GWDT based watershed fragment with which they had maximum overlap (Fig. 3(h)). To capture the nuclei embedded in bigger clusters (and missed by the previous steps), such clusters were identified using a two-dimensional feature (size more than 10,000 pixels and perimeter-to-area ratio (P2A) more than 1.4) classification system and then watershed was performed on each identified cluster. Due

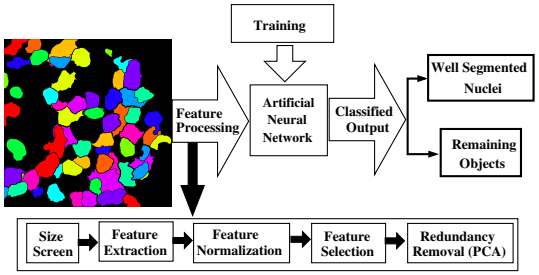


Fig. 4. Pattern Analysis Module

to the nuclei size variations from one dataset to the other, the watershed algorithms failed to identify potentially good nuclei by oversegmenting them. Hence, in the final step of the method, a tree based hierarchical merging strategy [12] was coupled with elliptical nuclear shape modeling to merge oversegmented nuclei. Fig. 3(i) shows the final output.

C. Pattern Classification Using ANN

The pattern analysis module is shown in Fig. 4. A 64 dimensional feature set was measured by augmenting the 24 dimensional feature set reported previously [4] with 40 new features to capture most of the pertinent morphological and textural properties of nuclei. The feature set was first normalized and then reduced prior to classification with an ANN which was used in place of an earlier stacked classifier [4] in order to improve the classification performance. Feature normalization plays a vital role in ANNs and is essential for numerical stability, hence, we tested 6 normalization techniques [13], namely, linear scaling to unit range, **Z**-Score, linear scaling to unit variance, transformation to uniform distribution, rank normalization and no normalization, to identify the best. Next, dependency ranking [14] was used to identify the most relevant features from the 64 dimensional feature set. It was calculated using,

$$D(i) = p(x_i, y) log \left| \frac{p(x_i, y)}{p(x_i)p(y)} \right| dx_i dy, \qquad (1)$$

where $D(i)$ is the dependency ranking score, x_i is the i^{th} element of the feature set and y is the set of output labels. Probability densities $p(\cdot)$ and joint probability densities $p(\cdot, \cdot)$ were calculated using histogram count. This was followed by a principal component analysis (PCA) based redundancy removal.

The ANN architecture had a single tansigmoidal hidden layer and a linear output layer. The classification problem was posed as a 2 class problem with output classes: 'Well Segmented Nuclei' and 'Remaining Objects' (i.e. poorly segmented nuclei). The ANN training set composed of feature vectors extracted from manually classified nuclei in 45 training images from the 9 datasets (D1–D9). 3 training methods were tested namely Levenberg-Marquardt Backpropagation Training, Conjugate gradient backpropagation with Powell-Beale restarts and Resilient backpropagation [15], all showing very similar results in terms of time and performance. The ANN based classification scheme was

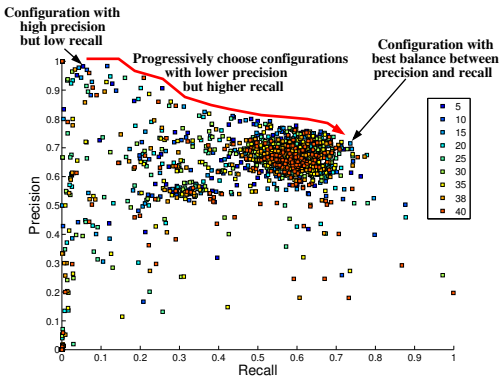


Fig. 5. Precision-Recall plot of 1620 configurations color coded with ANN hidden layer neuron count

used after the training phase (on 25% of data). The actual classification of the test set into good segmentation and inaccurate segmentation was entirely automatic.

III. Experiments and Results

The 9 datasets had in all 178 images and 1496 nuclei were manually identified as 'Well Segmented'. The 45 images used as the training set had a 'Well Segmented' manual nuclei count of 386. Though the segmentation algorithm segmented a reasonably high proportion of the imaged nuclei (compare Fig. 3(a) and 3(i)), the goal of the study was not to identify as many nuclei as possible, rather it was to automatically identify the subset of nuclei that were precisely segmented with high confidence.

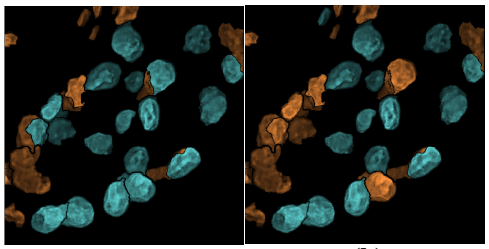
Experiments to identify the best possible configuration for 'Well Segmented' nuclei classification involved testing 1620 configurations of the pattern analysis module by varying the hidden layer neuron count, the normalization method, number of PCA dimensions and the number of features selected using dependency ranking. Fig. 5 shows the precision-recall plot for the configurations color coded with the hidden layer neuron count. The precision recall performance evaluation was performed on 133 images that did not belong to the 45 image training set. The best possible single configuration was identified as the one closest to the point $(1, 1)$ in the precision-recall plot having precision = 71.5% and recall = 73.6%. Fig. 6(a) shows segmented nuclei manually annotated as 'Well Segmented' (cyan) versus 'Remaining Objects' (orange) and Fig. 6(b) shows the automatically selected nuclei. Using this configuration the yield on the entire dataset was 1435 nuclei. Performance of the analysis pipeline was compared to that of a stacked classifier based classification system [4]. Table 1 shows the comparison between the stacked classifier and various configurations of the ANN classifier. The performance of the ANN based system was superior to that of the stacked classifier.

An unique and novel advantage of the new pattern recognition engine (PRE) lies in the flexibility to progressively select different configurations of the PRE (indicated by the red arrow in Fig. 5) in order to select the objects with the highest precision first. For instance, one can select an initial

TABLE I

TABLE SHOWING PERFORMANCE COMPARISON OF STACKED CLASSIFIER AND ANN CLASSIFIER SYSTEMS

Performance	Stacked Classifier	ANN Classifier (6 Configurations)					
Recall	63.89%	5.4%	20.9%	24.63%	53.09%	59.76%	73.6%
Precision	67.11%	98.2%	94.3%	85.96%	79.21%	77.68%	71.5%



(a) **(b)**

Fig. 6. (a) Hand annotated nuclei (Cyan - Well Segmented Nuclei, Orange - Remaining objects) (b) ANN selected nuclei

PRE configuration with a very high precision (98.21%), but with low recall (5.4%) (refer Fig. 5) to select a few nuclei. Then, another configuration, albeit at a slightly lower precision, can be used to select a few more nuclei. This process can be repeated until a sufficient number of nuclei have been automatically selected. Overall the precision will be significantly greater than 71.5% that was achieved for the single best configuration. This way one can be highly selective about the quality of the nuclei while accumulating sufficient nuclei for further analysis.

IV. CONCLUSIONS AND FUTURE WORK

We have described an automatic and intelligent image analysis pipeline that segments a high proportion of nuclei in tissue images and then screens out a certain number of nuclei with an acceptable degree of confidence about their segmentation accuracy. Although 3-D Z-stacks were acquired, initial data exploration revealed that the segmentation improvements offered by 3D analysis of the data would get outweighed by the adverse effects of significantly increased computational complexity, hampering time-efficient analysis of several hundred nuclei which improves the accuracy of the statistical analysis. Using a multistage watershed segmentation algorithm with superior segmentation performance, more features to identify well segmented nuclei and an ANN, the proposed methodology improves the nuclear screening efficacy of a previously reported stacked classifier based system. The proposed methodology speeds up the screening procedure by many folds, thus, enabling it to be used as a part of high throughput analysis. The method can be used to analyze nuclear features such as gene positioning or morphometric analysis to answer important questions in genome biology.

Some of the future work includes combining the segmentation results from the multistage watershed algorithms described here with hybrid level set watershed algorithm used previously [4], analysis of feature contributions, comparison of the ANN classifier to other methods such support vector machines and including features of the context around each segmented object in the selection process and progressively reconfiguring the classifier so that best segmented nuclei are selected first. Inspite of the problem of increased computational complexity, 3D analysis of the data using computationally efficient algorithms is also envisioned as a potential future work.

V. ACKNOWLEGMENTS

Fluorescence imaging was performed at the Fluorescence Imaging Facility, National Cancer Institute, NIH, Bethesda, MD.

REFERENCES

[1] K. J. Meaburn and T. Misteli, Locus-specific and activity-independent gene repositioning during early tumorigenesis, *The Journal of Cell Biology*, Vol. 180, pp 39-50, 2008

[2] K. J. Meaburn, P. R. Gudla, S. Khan, S. J. Lockett, and T. Misteli, Disease-specific gene repositioning in breast cancer, *The Journal of Cell Biology*, Vol. 187, no. 6, pp 801-812, 2009

[3] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, Second Edition, 2000, Wiley-Interscience, New York, NY

[4] K. Nandy, P. R, Gudla, K. J. Meaburn, T. Mistelli and S. J. Lockett, Automatic Nuclei Segmentation and Spatial FISH Analysis for Cancer Detection, *proceedings of 31st Annual International IEEE EMBS Conference*, Minneapolis, 2009, pp 6718-6721

[5] P. R. Gudla, K. Nandy, J. Collins, K. J. Meaburn, T. Misteli and S. J. Lockett, A High-Throughput System for Segmenting Nuclei Using Multiscale Techniques, *Cytometry A*, 2008, 73A, pp 451-466

[6] D. McCullough, P. Gudla, B. Harris, J. Collins, K. Meaburn, M. Nakaya, T. Yamaguchi, T. Misteli, S. J. Lockett, Segmentation of Whole Cells and Cell Nuclei From 3-D Optical Microscope Images Using Dynamic Programming, *IEEE Transactions on Medical Imaging*, 2008, 27, pp 723-734

[7] G. Lin, U. Adiga, K. Olson, J. F. Guzowski, C. A. Barnes and B. Roysam, A hybrid 3D watershed algorithm incorporating gradient cues and object models for automatic segmentation of nuclei in confocal image stacks, *Cytometry A*, 2003, 56A, pp 23-36

[8] V. Laurain, H. Ramoser, C. Nowak, G.E. Steiner and R. Ecker, "Fast Automatic Segmentation of Nuclei in Microscopy Images of Tissue Sections", *Proceedings of the 2005 IEEE Engineering in Medicine and Biology Conference*, Shanghai, China, 2005, pp 3367-3370

[9] *LastWave*, http://www.cmap.polytechnique.fr/˜bacry/LastWave

[10] F. Meyer, Topographic distance and watershed lines, *Signal Processing*, 38, pp. 113-125, 1994

[11] P. Soille, Morphological Image Analysis: Principles and Applications, *Springer-Verlag*, 1999

[12] G. Lin, M. K. Chawla, K. Olson, J.F. Guzowski, C.A. Barnes and B. Roysam, Hierarchical, Model-Based Merging of Multiple Fragments for Improved Three-Dimensional Segmentation of Nuclei, *Cytometry A*, 63A, pp 20-33, 2005

[13] S. Aksoy and R. Haralick. Feature Normalization and Likelihood-based Similarity Measures for Image Retrieval, *Pattern Recognition Letters*. 2000, 22, pp 563-582.

[14] I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection, *J Machine Learning Research*, 2003, 3, pp 1157-1182.

[15] Neural Network Toolbox, *Matlab 2008a*, The MathWorks, Inc., Natick MA