A caGRID-ENABLED, LEARNING BASED IMAGE SEGMENTATION METHOD FOR HISTOPATHOLOGY SPECIMENS

David J. Foran[§], Lin Yang[§], Oncel Tuzel[†], Wenjin Chen[§], Jun Hu[§], Tahsin M. Kurc[‡], Renato Ferreira[♯], Joel H. Saltz[‡]*

[§]The Cancer Institute of New Jersey, UMDNJ-RWJMS, Piscataway, NJ 08854
[†]Department of Computer Science, Rutgers University, Piscataway, NJ 08854
[‡]Center for Comprehensive Informatics, Emory University, Atlanta, GA 30322
[♯]Department of Biomedical Informatics, Ohio State University, Columbus, OH 43201

ABSTRACT

Accurate segmentation of tissue microarrays is a challenging topic because of some of the similarities exhibited by normal tissue and tumor regions. Processing speed is another consideration when dealing with imaged tissue microarrays as each microscopic slide may contain hundreds of digitized tissue discs. In this paper, a fast and accurate image segmentation algorithm is presented. Both a whole disc delineation algorithm and a learning based tumor region segmentation approach which utilizes multiple scale texton histograms are introduced. The algorithm is completely automatic and computationally efficient. The mean pixel-wise segmentation accuracy is about 90%. It requires about 1 second for whole disc (1024×1024 pixels) segmentation and less than 5 seconds for segmenting tumor regions. In order to enable remote access to the algorithm and collaborative studies, an analytical service is implemented using the caGrid infrastructure. This service wraps the algorithm and provides interfaces for remote clients to submit images for analysis and retrieve analysis results.

Index Terms— Segmentation, Tissue Image Analysis

1. INTRODUCTION

Breast cancer accounts for about 30% of all cancers and 15% of all cancer deaths in women in the United States. Current therapies and treatment regimens are based upon classification strategies which are limited in terms of their capacity to identify specific tumor groups exhibiting different clinical and biological profiles. Tissue microarray (TMA) technique enables investigators to extract small cylinders of tissue from histological sections and arrange them in a matrix configuration on a recipient paraffin block such that hundreds can be analyzed simultaneously [1, 2]. An alternate, but less utilized approach is to sequentially digitize each specimen for subsequent semi-quantitative assessment [3]. Both strategies ultimately involve the interactive evaluation of TMA samples which is a slow, tedious process that is prone to error. Processing the specimen using a reliable, image-based analysis system could reduce the cost and patient morbidity. There has been increasing interest in investigating the image analysis algorithm for digitized TMA tissue microarray images [4]. However, to our knowledge, most of these studies run as stand-alone programs, which limits the scale and throughput as a result of the computational complexity required by many of the algorithms used for analysis.

The biomedical research community has recognized the importance of collaborative use of databases and analysis systems, developed by independent research groups and/or hosted by different institutions, in order to target complex diseases. There are several large scale projects, driven by community needs, that develop tools and infrastructure to support federation of information and analytical resources for basic, clinical, and translational research. An example in the cancer research field is the cancer Biomedical Informatics Grid (caBIG®, http://cabig.nci.nih.gov) program, sponsored by the National Cancer Institute. The goal of this program is to develop informatics standards, a common suite of applications, and a Grid infrastructure to assist more effective sharing of data and analytical resources across institutions and support coordinated multi-institutional projects. The CardioVascular Research Grid (CVRG, http://cvrgrid.org) is another example. The CVRG is developing a suite of tools, applications, and a federated infrastructure (building on the caBIG® caGrid architecture[5]) to support information sharing and collaborative studies in the cardiovascular research community.

In this paper, we describe a learning based segmentation algorithm for analyzing digitized breast tissue specimens. In Section 2 we present the details of the segmentation algorithm. The implementation using caGrid for remote access and collaboration is described in Section 3. Section 4 provides the experimental results and Section 5 concludes the paper.

2. SEGMENTATION

Image segmentation is the process of delineating an image into "homogeneous" regions based on the similarity of pixel attributes. In our applications, the pre-processing step of

*This research was funded, in part, by grants from the NIH through contract 5R01EB003587-03, and through NHLBI R24-HL085343, NIH R01-LM009239, NCI contract no. 79077CBS10, and NCI contract no. N01-CO-12400; from the National Institute of Biomedical Imaging and Bioengineering and contract 5R01LM009239-02 from the National Library of Medicine. Additional funds were provided by the Department of Defense via grant number W81XWH-06-1-0514 and National Science Foundation grant CNS-0615155.

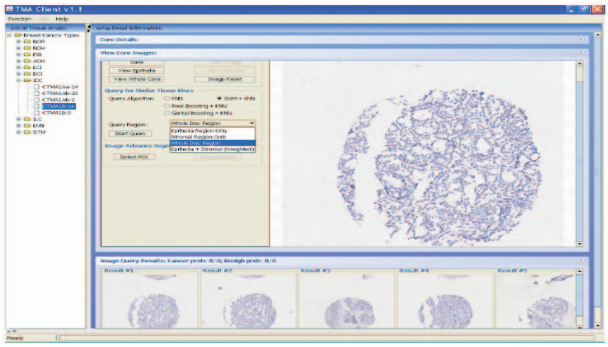

Fig. 1. *TMA-Miner* Prototype Client Interface.

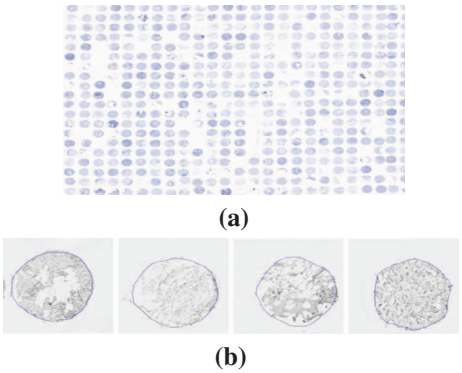

Fig. 2. The classification results. (a) A sample imaged tissue microarray. (b) The delineation of the outer contours of a few representative breast tissue discs.

the segmentation algorithm automatically detects the outer contour of imaged tissue discs. An Adaboost classifier was trained using a multiscale texton histogram as the feature vector. The whole procedure has been implemented as a caGrid analytical service and can be launched remotely by remote clients. The Java based image analysis interface is shown in Figure 1.

2.1. Whole Disc Delineation

The algorithm begins by finding the outer contour of each whole breast tissue disc. This is achieved by first applying a simple adaptive threshold to provide a binary mask for the tissue disc. The algorithm then roughly estimates the outer boundary of the binary disc as the region of interest (ROI). Given a set of point set S, with points $p_1, ..., p_N$, the outer envelop is given by the following equation:

$$H = \left\{ \sum_{j-1}^{N} \lambda_j p_j : \lambda_j \geqslant 0 \text{ for all } j \text{ and } \sum_{j-1}^{N} \lambda_j = 1 \right\}. \quad (1)$$

A deformable model [6] is further applied to extract the breast cancer disc from the background.

In Figure 2, we show the unsupervised delineation of the outer boundary of a few representative tissue discs. After segmenting the whole disc mask, the segmentation of the ROI is performed using texton histograms.

2.2. Learning Based Tumor Region Segmentation

Textons [7] are defined as repetitive local features that humans perceive as being discriminative between textures. We use the multiple scale Schmid filter bank [8] composed of 13 rotation invariant filters:

$$F(r, \sigma, \tau) = F_0(\sigma, \tau) + \cos\left(\frac{\pi r \tau}{\sigma}\right) e^{-\frac{r^2}{2\sigma^2}} \quad (2)$$

The image filtering responses are clustered using K-means to generate a large code book. A texton library is constructed from the corresponding cluster centers. The pixel-wise segmentation of imaged breast tissue is performed by classification. Based on the labeled ground truth masks, 2000 positive and negative pixels are extracted from the tumor and non-tumor regions in the image. The appearance of the neighbors of each training pixel is modeled by a compact quantized description - texton histogram, where each pixel is assigned to its closest texton using the following equation:

$$h(i) = \sum_{j \in I} count(T(j) = i) \quad (3)$$

Here I denotes breast tissue image, i is the i-th element of the texton dictionary, and $T(j)$ returns the texton assigned to pixel j. The windowed texton histogram is computed around each individual training pixel.

After normalization, the texton histogram actually represents the texton channel frequency distribution in a local neighborhood around the centered pixel. In order to compensate for scale changes, the texton histogram is extracted from 5 different window sizes (4, 8, 16, 32, 64 pixels, respectively) and concatenated into one large feature vector. This concatenated texton histogram is used as features to train the classifiers. The integral histogram [9] is used to calculate the windowed texton histogram. The algorithm starts by exploiting the spatial arrangement of data points. It then recursively propagates an aggregated histogram. The aggregated histogram starts from the origin and traverses through the remaining points along a scan-line. At each step, a single bin is updated using the values of the integral histogram at the previous visited neighboring data points. The integral histogram method speeds up feature extraction significantly.

The Adaboost is chosen as the classifier for segmentation. AdaBoost works by sequentially applying a classification algorithm on a reweighed version of the training data and produces a sequence of weak classifiers. The weak leaner used in our experiments is classification stump. The strong classifier is assembled from all the weak classifiers to minimize the cost function representing the classification accuracy. Given a test

Input: Given the input imaged tissue specimen.
Training:

- Extract the multiple-scale texton histogram as the input features $\mathbf{x}$.
- Initialize the weights $w_i = 1/n, i = 1, ..., n$. Set $b(\mathbf{x}) = 0$.
- For $j = 1...J$
  - Each training sample is assigned its weight w_i. Find the decision stump $h_j(\mathbf{x})$ which minimizes the total weighted classification errors.
  - Update using $b(\mathbf{x}) = b(\mathbf{x}) + h_j(\mathbf{x})$.
  - $\alpha_j = \frac{1}{2} \ln \frac{1-\epsilon_j}{\epsilon_j}$.
  - Update the weights $w_i = w_i e^{-y_i \alpha_j h_j(\mathbf{x})}$ and renormalize w_i.
  - Save the j-th decision stump $h_j(\mathbf{x})$.

Testing:

- Apply the adaptive threshold
- Calculate the initial position by finding the outer pixels
- Apply deformable model to delineate the outer contour
- Construct the multiple-scale texton histogram using the same procedure in training stage
- Utilize the trained classifier to provide the output label for each pixel: $sign\left[b(\mathbf{x})\right] = sign\left[\sum_{j=1}^{J} h_j(\mathbf{x})\right]$.

Fig. 3. The segmentation procedure which applied deformable model to find outer contour and multiple scale texton histogram for tumor region segmentation.

image, we apply the trained strong classifiers for each pixel and separate the image into tumor and non-tumor regions.

Using the multiscale texton histogram, integral histogram and AdaBoost, a fast and accurate pixelwise segmentation algorithm can be implemented for delineating the tumor region in an imaged breast cancer specimen. The pseudocode of the algorithm is shown in Figure 3

3. IMPLEMENTATION OF GRID SERVICES FOR REMOTE ACCESS AND COLLABORATION

One of our goals in this project is to facilitate remote access to analysis methods and analysis results by researchers and among collaborating teams and to enable efficient execution of expensive analysis on high-end systems. We employ Grid computing and high performance computing frameworks for this purpose. In this work we have adopted a service oriented implementation to support remote access to analysis programs. This implementation encapsulates an analysis method or application as a service. The analysis application's functionality is accessed remotely and programmatically through application-specific service interfaces. With a service oriented design and implementation, a heterogeneous collection of analysis programs (which may be implemented as Matlab scripts, Java codes, or C++ programs) can be accessed through well-defined and published interfaces. This facilitates more effective and easier federation of multiple analytic resources in a collaborative environment. Moreover, the backend analysis program can be deployed on a parallel machine for faster execution of requests without requiring modifications to client programs. We use the caGrid infrastructure [5] for Grid-enabled deployment of our analysis methods. caGrid is the core Grid architecture of caBIG®. It is implemented as a service oriented architecture with extended support on service metadata, interoperability through published XML schemas and common data elements, and security.

Our choice of caGrid as the underlying infrastructure is motivated by several factors. First, caGrid is employed by both the caBIG program and the CVRG project. Implementing our Grid services using caGrid would enable us to interoperate with tools and resources developed by those communities. Second, caGrid provides higher level tools and core services such as Introduce [10] for service development and deployment and GAARDS [11] for security support on top of low level Grid middleware. These tools make it easier to develop and deploy interoperable services and implement Grid-enabled authentication and authorization support for a service. Third, a service oriented system provides flexibility in organizing and combining the steps of an analysis process into services. For instance, each step may be implemented as a separate service or multiple steps can be combined into one service.

We have developed a suite of services for analysis of TMA data. One implementation treats each step in the analysis process as a separate service. The advantage of this approach is that the client can compose different analysis processes using a subset of these services. The client can also replace a service (a step) with another semantically equivalent service, which may be implementing a different algorithmic variation of the analysis step. The disadvantage is that it introduces overheads because of multiple service invocations and because data is exchanged through service interfaces, rather than using native data formats and file or memory copies. A more recent implementation combines multiple steps (each of which is implemented as a stand-alone program) into a single caGrid analytical service. This implementation has less overhead, but offers less flexibility to clients. The implementation of this service has been done using the caGrid Introduce toolkit. We have implemented a service interface and skeleton using Introduce. The service interface accepts a TMA disc image and input parameters used by the analysis programs. It returns a texton histogram as the analysis result. When an image is received by the service, the service stores the image into a file and invokes the backend analysis programs passing them the image file. Once the analysis of the image has been completed, the service converts the results into an object, which

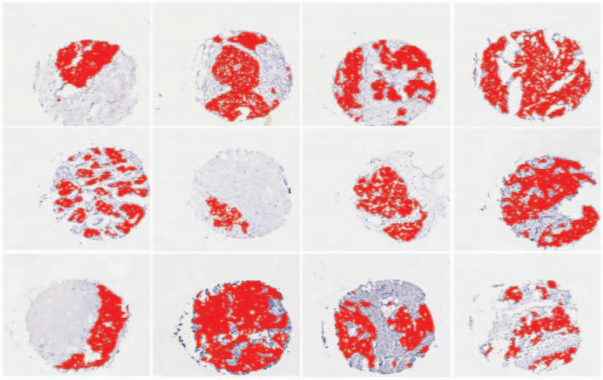

Fig. 4. The representative segmentation results . The segmented mask is overlayed on the original images.

represents the texton histogram, and returns the object to the client. We are in the process of extending this service to accept a collection of images and make use of a parallel machine so that multiple images can be processed concurrently reducing the overall execution time.

4. EXPERIMENTAL RESULTS

The tissue microarrays used in our experiments were prepared by different institutes: the Cancer Institute of New Jersey, Yale University, University of Pennsylvania and Imgenex Corporation, San Diego, CA. Diaminobenzidine (DAB) and hematoxylin were used to stain the tissue samples. To date over 300 immunostained microscopic specimens, each containing hundreds of tissue image, were digitized at 40 volume scan using the Trestle/Zeiss MedMicro, the whole slide scanner system. The output images typically contain a few billions of pixels and are stored as a compressed tiled TIFF file sized at about two gigabytes.

We obtained 100 breast cancer specimens for which ground truth tumor masks were hand-drawn by a board-certified anatomic pathologist. Compared with the doctor's annotation, the algorithm provided a pixel-wise segmentation accuracy around 90% with the average false positive rate 6.62% and the average false negative rate 3.15%. Some of the segmentation results are shown in Figure 4. The algorithm is implemented using C++ and computationally efficient. On a PC with Duo Core Processor 1.8GHz and 2G memory, the whole disc delineation took only 1 second for a 1024*1024 images, while the segmentation of the tumor region took less than 5 seconds.

5. CONCLUSIONS

In this paper, we have presented a robust, fast and accurate segmentation algorithm for digitized tissue microarray images. A novel aspect of this algorithm is that instead of building specific models of the specific problem, all the major steps in the segmentation process are based on learning. This characteristic of the algorithm makes it possible to extend the algorithm to other types of digitized pathology specimen segmentation. Our implementation leverages emerging service oriented Grid architectures for remote access to the algorithm for collaborative studies. We believe the availability of extensible algorithms deployed as services has tremendous potential to significantly improve scientific research that makes use of biomedical imaging.

6. REFERENCES

[1] J. Kononen, L. Bubendorf, A. Kallioniemi, M. Barlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter, and O. P. Kallioniemi, "Tissue microarrays for high-throughput molecular profiling of tumor specimens," *Nat Med*, vol. 4, no. 7, pp. 844–847, 1998.

[2] D. L. Rimm, R. L. Camp, L. A. Charette, J. Costa, D. A. Olsen, and M. Reiss, "Tissue microarray: A new technology for amplification of tissue resources," *Cancer Journal*, vol. 7, no. 1, pp. 24–31, 2001.

[3] N. R. Mucci, G. Akdas, S. Manely, and M. Rubin, "Neuroendocrine expression in metastatic prostate cancer: Evaluation of high throughput tissue microarrays to detect heterogeneous protein expression," *Human Pathology*, vol. 31, no. 4, pp. 406–414, 2000.

[4] Thomas J. Fuchs, Peter J. Wild, Holger Moch, and Joachim M. Buhmann, "Computational pathology analysis of tissue microarrys predicts survival of renal clear cell carcinoma patients," *MICCAI*, vol. 5242, pp. 1–8, 2008.

[5] S. Oster, S. Langella, S. Hastings, D. Ervin, R. Madduri, J. Phillips, T. Kurc, F. Siebenlist, P. Covitz, K. Shanbhag, I. Foster, and J. Saltz, "caGrid 1.0: An Enterprise Grid Infrastructure for Biomedical Research," *Journal of the American Medical Informatics Association (JAMIA)*, vol. 15, pp. 138–149, 2008.

[6] L. Yang, P. Meer, and D.J. Foran, "Unsupervised segmentation based on robust estimation and color active contour models," *IEEE Trans. on Information Technology in Biomedicine*, vol. 9, pp. 475–486, 2005.

[7] B. Julesz, "Textons, the elements of texture perception, and their interactions.," *Nature*, vol. 290, no. 5802, pp. 91–97, 1981.

[8] C. Schmid, "Constructing models for content-based image retrieval," *CVPR*, vol. 2, pp. 39–45, 2001.

[9] F.M. Porikli, "Integral histogram: A fast way to extract histograms in cartesian spaces," *CVPR*, vol. 1, pp. 829–836, 2005.

[10] S. Hastings, S. Oster, S. Langella, D. Ervin, T. Kurc, and J. Saltz, "Introduce: An open source toolkit for rapid development of strongly typed grid services," *Journal of Grid Computing*, vol. 5, no. 4, pp. 407–427, 2007.

[11] S. Langella, S. Hastings, S. Oster, T. Pan, A. Sharma, J. Permar, D. Ervin, B. Cambazoglu, T. Kurc, and J. Saltz, "Sharing data and analytical resources securely in a biomedical research grid environment," *Journal of American Medical Informatics Association*, vol. 15, no. 3, pp. 363–373, 2008.