Ovarian Cancer Detection using Hierarchical Normalized Cuts

R. Reeja and A. Ann.Romalt

Abstract--- *Ovarian cancer is cancerous growth arising from the ovary. The ovaries are part of a woman's reproductive system. In this , it is demonstrated that an algorithm which is flexible, robust, accurate, termed hierarchical normalized cuts (HNCuts) for the specific problem of quantifying extent of vascular staining on ovarian cancer tissue microarrays. The high efficiency of HNCut is driven by the use of a hierarchically represented data structure that allows us to merge two powerful image segmentation algorithms. The HNCuts combines two powerful image segmentation algorithms named frequency weighted mean shift algorithm and the normalized cuts algorithm. HNCuts rapidly traverses a hierarchical pyramid, generated from the input image at various color resolutions, enabling the rapid analysis of large images. HNCut is easily generalizable to other problem domains and only requires specification of a few representative pixels (swatch) from the object of interest in order to segment the target class. The success in accurately quantifying the extent of vascular stain on ovarian cancer TMAs suggests that HNCut could be a very powerful tool in digital pathology and bioinformatics applications where it could be used to facilitate computer-assisted prognostic predictions of disease outcome.*

Keywords--- *Hierarchical Normalized Cuts (HNCuts), Mean Shift, Normalized Cuts, Segmentation, Tissue Micro Array (TMA)*

I. INTRODUCTION

BY the introduction of whole slide digital scanners, histological data have now become amenable to digital and quantitative image analysis . It is reported that 21 990 women will be diagnosed with and 15460 women will die of cancer of the ovary (OCa) in 2011. The five-year survival rates of these women are highly correlated with the early detection of OCa. Recent work [4] suggests that specific tumor vascular biomarkers (TVMs), identifiable on OCa TMAs, could have prognostic significance, which would enable not only predicting the aggressiveness of the disease, but could also help in tailoring a personalized treatment regime for the patient.

The manual quantification of extent of biomarker staining is, however, a laborious, time consuming, and error-prone affair. Consequently, there is a real need for high-throughput quantitative image analysis algorithms which can automatically and efficiently estimate biomarker extent on very large pathology slides in a few seconds.

Most previous computerized image analysis algorithms for TMAs have involved thresholding-based schemes. These methods are known to be highly sensitive to even slight changes in color and illumination. Clustering based approaches, including k-means, have also been investigated for the analysis of TMAs. However, k-means is a nondeter-ministic algorithm and is highly sensitive to the initial choice of cluster centers. Active contour schemes, while suitable for cell and nuclear segmentation in digital pathology, are not ideally suited to the problem of pixel-level classification. Additionally, they are typically infeasible for problems where hundreds of objects need to be concurrently segmented on very large images. It was shown that using hierarchical normalized cuts (HNCuts) as a preprocessing step for an active contour approach drastically improved computation time of their active contour approach by providing a significantly better initial estimate to the region.

Normalized Cuts (NCuts) is among the final mature descendants from a series of graph-cutting techniques ranging from max cut to min cut. It is a popular scheme in spite of its main drawbacks: 1) the large number of calculations needed for determining the affinity matrix and 2) the time-consuming eigen value computation. For large images, the computation and overhead of these border on the infeasible. Consequently, a significant amount of research has focused on avoiding their direct calculations.

The mean-shift algorithm (MS) has been employed and modified in as an unsupervised technique for mode discovery instead of k-means. The MS algorithm attempts to identify the cluster mean within a predefined bandwidth. By using a steepest gradient approach, a fast convergence to the set of true means of the statistical data can be found. The improved fast Gauss transform (IFGT) implementation of the MS algorithm allowed computation times for large images to become reasonable.

There are two major differences between this study and a previous, preliminary conference version of this paper. The first difference is that this study uses a frequency weighted mean shift (FWMS) which significantly improves the computation time over. The second is an extension of the original work by the inclusion of a number of additional experiments to rigorously and quantitatively evaluate our scheme on a much larger data cohort compared to what was initially presented in. The strength of HNCut is in that it combines a powerful unsupervised clustering technique with an equally powerful graph partitioning scheme. By performing clustering and partitioning in the color space, the HNCut algorithm is highly efficient and precise. For large images, such as TMAs where

R. Reeja, Student, M.E. Computer Science and Engineering, Udaya School of Engineering, Tamil Nadu, India. E-mail: reeja.thoppil@gmail.com

A. Ann.Romalt, M.E., Assistant Professor, Computer Science and Engineering, Udaya School of Engineering, Tamil Nadu, India.

ISBN 978-93-82338-35-2 | © 2013 Bonfring

there are often many fewer unique colors than pixels, performing the analysis in the color as opposed to the spatial domain could result in significant improvements in computational processing time. HNCut only requires specifying a few representative pixels from the target class and, unlike more traditional supervised classification algorithms, does not require more detailed target object annotation. More importantly, the HNCut algorithm is more flexible compared to supervised schemes in its ability to segment different object classes. The combination of both the high-throughput efficiency and flexibility of HNCut makes it ideally suited to applications requiring high-throughput analysis, such as quantifying the expression of biomarkers on TMAs. In this paper, we demonstrate the specific application of HNCut to a problem of automated quantification of stain extent associated with a vascular marker on OCa TMAs. The rest of this paper is organized as follows.

The work presented in this paper represents important methodological and clinical contributions summarized as follows.

1. A new minimally supervised hierarchical segmentation approach that combines a FWMS and normalized cuts (HNCuts) for pixel-level detection and classification. HNCut is able to segment very large images rapidly.
2. HNCut is largely insensitive to choice of parameter valueand is able to discriminate between regions with similar color values. The parameters for NCuts are automatically computed, and the parameters for the FWMSareautomatically adjusted based on the variance of the output.
3. This study represents the first attempt, to our knowledge, to accurately quantify a vascular marker on OCa TMAs with the ultimate objective of creating a quantitative image based metric for OCa prognosis and survival.

II. Related Work

The attempt to merge NCuts and mean shift is not new . To overcome the computational issues associated with NCut, a novel approach of combining both the MS and NCut algorithms was presented. Clustering the image by running the MS algorithm to convergence produced class assignments for the pixels. By taking the average intensity value of the regions obtained via the MS clustering step and using them as the vertices in the NCut algorithm, a significant speed improvement was obtained. It was later noticed that when points of similar values are within an neighborhood of each other, their contribution to the overall system can be merged, providing an efficiency improvement by reducing the number of computations needed per iteration. We use this to extend the MS work in a hierarchical fashion which is more pertinent and amenable to problems in digital pathology and biomedical imaging. This allows us to perform the same detection or segmentation task in half the time. The proposed algorithm is specifically designed for rapid extraction of pixels of interest in a minimally supervised manner, as opposed to unsupervised clustering which is insensitive to the user's domain knowledge as the aforementioned approaches take.

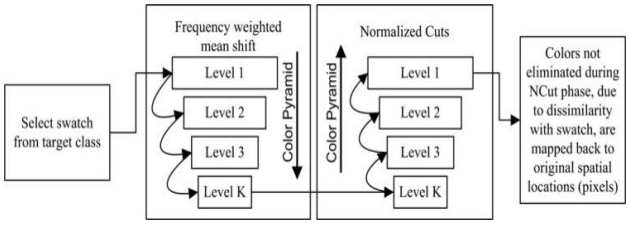

Figure 1: Flow Chart of the HNCut Process

Proceeding left to right, the user selects the domain swatch, followed by the FWMS of the image. This results in the original image being decomposed into multiple levels of color resolution, which is then followed by the application of NCut at each of the color resolutions generated. At each pyramid level, colors not deemed to be part of the swatch are eliminated. Following the application of NCut on the color pyramid (from the lowest to the highest color resolution), the color values that have not been eliminated are mapped back to the spatial domain via their original pixel locations, and the final segmentation is obtained.

First manually identify the desired target class based on individual representative colors selected from the target class by a user. This swatch, which can be changed based on the desired target class or domain, lends HNCut significant flexibility and ease of use. Second, to our knowledge, this is the first attempt at combining an FWMS with a partitioning algorithm that accomplishes the same task as MS but does it significantly faster. The FWMS exploits the fact that as each iteration of MS completes, more points converge. In this paper how the convergence of our novel FWMS scheme allows us to perform clustering 15 times faster than the traditional MS algorithm.

III. Description of HNCut

3.1 Overview

A high-level overview of the four stages associated with the HNCut algorithm. Each of these stages are discussed in detail in the following sections. We present an overview here to guide the reader through the various stages.

We start by requiring the user to select a few sample pixels from the target class from an image. We use these pixels to guide the subsequent pixel classification process across all images in the same domain. Next, we employ the MS algorithm on the color values in the image to form a hierarchical data structure (represented by the levels in the color pyramid in the second box in Fig. 2). Intuitively, the FWMS algorithm allows for identification of color values which are within some specified tolerance of each other and assigns them to the same mode. Employing the NCuts operation only on the unique values at each level of the pyramid, as opposed to all possible color values, allows for a factorization resulting in significantly fewer computations. An illustration of the application of the scheme to an OCa TMA, for detecting a TVM. We then compute the weight for each unique mode, which reflects the actual frequency of the number of pixels associated with it. Using this pyramid, we can drastically reduce the large segmentation problem in the color space to a set of much smaller graph partitioning problems (the third box from the left in Fig. 2),

ISBN 978-93-82338-35-2 | © 2013 Bonfring

which we show can be solved far more efficiently by NCut. By starting at the bottom of the pyramid, we partition the unique values (typically on the order of ten values) into two sets such that all of the values selected by the user in the first step are assigned to the first partition. Subsequently, we eliminate the second partition and map the colors in the first partition to an immediately higher color resolution level in the pyramid. This process continues until the entire pyramid is traversed. The last step involves mapping the color values not eliminated back into the spatial domain.

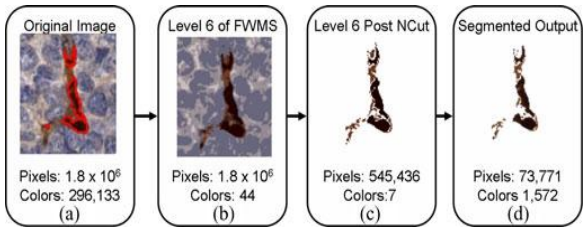

Figure 2 (a) Original image with desired TVM stain enclosed in red. (b) Image at the bottom of the color pyramid during FWMS. (c) Image at the bottom of the color pyramid following application of NCuts. (d) Final segmentation results obtained by mapping colors not eliminated by HNCut spatially onto the original image. Note that between (a) and (b), a significant reduction in color resolution occurs, which allows NCuts to be performed on an image with several orders of magnitude fewer colors compared to the original image (a). NCut is then applied at progressively higher color resolutions, while at each pyramid level, colors not deemed to be part of the swatch are eliminated. The colors retained at the highest resolution are then spatially mapped onto the corresponding pixels to yield the final segmentation

The hierarchical set of operations described previously makes for an extremely efficient and accurate algorithm; thus, applying the NCut at the lowest levels of the pyramid is relatively simple to do and encourages a more sophisticated definition of pixel affinity. While in this paper only chromatic information was leveraged, the method is easily and efficiently extensible to incorporate additional image features (e.g., texture).

Fig. 2 displays an image from our dataset undergoing the HNCut procedure, with the intent of quantification of the vascular marker stain (brown color). The numbers shown in the boxes in Fig. 3 represent the reduced number of colors and pixels generated by the HNCut scheme at different levels of the pyramid within a single cylinder (1500 × 1500 pixels, 300000 colors) from a TMA.

3.2 Notation

An image scene is defined as C = (C, f) where C is a 2-D Cartesian grid of N pixels, $c \in C$, where c = (x, y). f is a color intensity function, where $f \in R^3$.

We define as $F_1 \in R^3$ the vector of colors associated with all pixels $c \in C$ at the full color resolution (top of the color pyramid). The elements of F_1, namely $f_{1,i}$, are derived such that for pixel c_i , $f_{1,i} = f(c_i)$ and $f_{1,i} \in R^3$.

3.3 FWMS for Reducing the Number of Colors

The MS algorithm is used to detect modes in data using a density gradient estimation. By solving for when the density gradient is zero and the Hessian is negative semidefinite, we can identify local maxima. For a more detailed explanation of the algorithm, we refer the reader to [4]. Thereby avoiding the explicit calculation of $G(f_{k,j} - f_{k,\beta}2)$, where $j, \beta_1, \beta_2 \in \{1, \ldots, N\}$, $k \in \{1, \ldots, K\}$. This results in one less computation for the Gaussian, which is by far the most expensive operation in the entire MS clustering process.

3.4 NCUTS on FWMS Reduced Color Space

NCuts is a graph partitioning method, used to separate data into disjoint sets. For our problem, the hierarchical pyramid created by FWMS at various levels of color resolution (F1, F2,…..,Fk) serves as the initial input to the NCut algorithm. The NCut takes a connected graph G = (E, V), with vertices (V) and edges (E) and partitions the vertices into disjoint groups. By setting V equal to the set of color values Fk , and having the edges represent the similarity (or affinity)between the color values, we can separate the vertices into groups of similar color values. The NCut is defined as the process by which the removal of edges leads to two disjointed partitions A and B such that the variance of values (in our case colors) in A and B are minimized and the difference in average value (intensity of colors) between A and B is maximized.

IV. Experimental Setup

All experiments were performed after converting the RGB input images to the HSV color space. The FWMS was performed using $\sigma_{MS} = 0.05$. NCut was subsequently performed using the Silverman function to determine the value for the initial σ_{NCut}. When the number of remaining clusters fell below this value, it was reset to the square root of the number of remaining clusters. Since the human visual system is unable to easily discriminate between subtle variations of the same color, we can set to a relatively large value. The easiest way to apply this requirement in an algorithmic form is to simply choose the desired precision level and then simply round the value to the right of that place. The subsequent procedure of locating unique values and computing their frequencies is as simple as generating a histogram of the data values with each unique value occupying its own bin. This is a significant benefit, as the production of histograms is not only well studied but easily transformable into a parallel computing problem.

We compared the detection performance of HNCut with *k*-means. A standard *k*-means algorithm was performed using ten clusters. Since *k*-means is not deterministic and is notoriously sensitive to the choice of cluster centers, offline experiments were performed to identify initial cluster centers which were qualitatively determined as being optimal. It is also worth noting that *k*-means does quite poorly. There is no variance associated with the algorithm since we determined the optimal centers offline, thus removing the nondeterministic aspect of the scheme.

ISBN 978-93-82338-35-2 | © 2013 Bonfring

V. Conclusion and Future Research

In this paper, I have presented a minimally supervised segmentation scheme termed HNCuts for precise and accurate quantification of extent of vascular staining of OCa TMAs. The extent and severity of this vascular stain has been predicted to be an important prognostic marker in predicting outcome of women with OCa. The strength of HNCut is derived from the fact that it integrates the best of both an FWMS clustering and the NCut algorithm. While other schemes have been previously proposed in an attempt to combine both mean shift and NCuts, we believe that HNCut is the only approach which provides the flexibility to be able to extract different target classes based on user input.

Additionally, the HNCut algorithm's hierarchical usage of the color space, a novel feature, allows it to operate faster compared to other similar approaches. By using our newly presented combination of FWMS and NCuts, and by operating in the color space, HNCut is able to handle large images efficiently. HNCut was found to be 62% faster compared to a state-of-the-art supervised classification scheme. A major advantage of HNCut, apart from its efficiency and accuracy, is that it is not encumbered by the need for precisely annotated training data.

References

[1] A. Madabhushi, "Digital pathology image analysis: Opportunities andchallenges," Imag. Med., Vol. 1, Pp. 7-10, 2009.

[2] G. Alexe, J. Monaco, S. Doyle, A. Basavanhally, A. Reddy, M. Seiler,S. Ganesan, G. Bhanot, and A. Madabhushi, "Towards improved cancer diagnosis and prognosis using analysis of gene expression data and computer aided imaging," Exp. Biol. Med., Vol. 234, Pp. 860-879, 2009.

[3] H. Rui and M. J. Le Baron, "Creating tissue microarrays by cutting-edgematrix assembly," Expert Rev. Med. Devices, Vol. 2, No. 6, Pp. 673-680, Nov. 2005.

[4] R. Buckanovich, D. Sasaroli, A. O'Brien-Jenkins, J. Botbyl, R. Hammond, D. Katsaros, R. Sandaltzopoulos, L. A. Liotta, P. A. Gimotty, andG. Coukos, "Tumor vascular proteins as biomarkers in ovarian cancer,"J. Clin. Oncol., Vol. 25, Pp. 852-861, Mar. 2007.

[5] H. Vrolijk, W. Sloos, W. Mesker, P. Franken, R. Fodde, H. Morreau, andH. Tanke, "Automated acquisition of stained tissue microarrays for highthroughput evaluation of molecular targets," J. Mol. Diagnost., Vol. 5, No. 3, Pp. 160-167, 2003.

[6] J. Wu, J. Dong, and H. Zhou, "Image quantification of high-throughputtissue microarray," in Proc. Soc. Photo-Opt. Instrum. Eng., Mar. 2006,Vol. 6143, Pp. 509520.

[7] A. Rabinovich, S. Krajewski, M. Krajewska, A. Shabaik, S. Hewitt,S. Belongie, J. Reed, and J. H. Price, "Framework for parsing, visualizing and scoring tissue microarray images," IEEE Trans. Inf. Technol. Biomed., Vol. 10, No. 2, Pp. 209-219, Apr. 2006.

[8] W. Zhong, G. Altun, R. Harrison, P.C. Tai, and Y. Pan, "Improvedk-means clustering algorithm for exploring local protein sequence motifs representing common structural property," IEEE Trans. Nano Biosci., Vol. 4, No. 3, Pp. 255-265, Sep. 2005.

[9] H. Fatakdawala, J. Xu, A. Basavanhally, G. Bhanot, S. Ganesan, M.Feldman, J. Tomaszewski, and 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] L. D. Cohen and R. Kimmel, "Global minimum for active contour models:A minimal path approach," IEEE Int. J. Comput. Vis., Vol. 24, No. 1, Pp. 57-78, Jan. 1997.

[11] J. Xu, A. Janowczyk, S. Chandran, and A. Madabhushi, "Aweighted mean shift, normalized cuts initialized color gradient basedgeodesic active contour model: Applications to histopathology image segmentation," Proc. SPIE, 7623, 76230Y, 2010, Available:http://dx.doi.org/10.1117/12.845602.

[12] Zhuowen Tu, "Probabilistic boosting-tree: Learning discriminative modelsfor classification, recognition, and clustering," in Proc. IEEE Int. Conf.Comput. Vis., Pp. 1589-1596, 2005.

[13] P. Tiwari, M. Rosen, G. Reed, J. Kurhanewicz, and A. Madabhushi, "Spectral embedding based probabilistic boosting tree (sceptre): Classifyinghigh dimensional heterogeneous biomedical data," in Proc. Med. Image Comput. Comput. Assist. Intervention, Vol. 1, Pp. 844-851, 2009.

[14] G. Carneiro, B. Georgescu, S. Good, and D. Comaniciu, "Detection andmeasurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree".

ISBN 978-93-82338-35-2 | © 2013 Bonfring