CONTENT BASED SUB-IMAGE RETRIEVAL SYSTEM FOR HIGH RESOLUTION PATHOLOGY IMAGES USING SALIENT INTEREST POINTS

Neville Mehta, Raja' S. Alomari, and Vipin Chaudhary

Department of Computer Science and Engineering
University at Buffalo, Buffalo, NY 14260

ABSTRACT

Content-based image retrieval systems for digital pathology require sub-image retrieval rather than the whole image retrieval for the system to be of clinical use. Digital pathology images are huge in size and thus the pathologist is interested in retrieving specific structures from the whole images in the database along with the previous diagnosis of the retrieved sub-image. We propose a content-based sub-image retrieval system (sCBIR) framework for high resolution digital pathology images. We utilize scale-invariant feature extraction and present an efficient and robust searching mechanism for indexing the images as well as for query execution of sub-image retrieval. We present a working sCBIR system and show results of testing our system on a set of queries for specific structures of interest for pathologists in clinical use. The outcomes of the sCBIR system are compared to manual search and there is an 80% match in the top five searches.

***Index Terms*—** sCBIR, Computer Aided Diagnosis, scale-invariant features, IHC, digital pathology.

1. INTRODUCTION

Digital anatomic pathology has been attracting many researchers over the last two decades. High resolution scanners allow remote pathology diagnosis and consulting. Furthermore, having digitized anatomic pathology slides allow building databases for aiding pathologists in diagnosis by querying similar previously diagnosed cases. However, high resolution databases demand high storage and indexing capabilities due to large sizes of these images [1].

Many Content-Based Image Retrieval (CBIR) systems have been built for various applications that target full image retrieval such as [2, 3] and few efforts on sub-image retrieval such as [4, 5].

Clinical and efficient usage of CBIR systems for digital pathology computer aided diagnosis (CAD) systems require sub-image retrieval for querying specific structures in the high resolution images along with the diagnosis as shown in Fig. 1. Pathologists are interested in querying about specific structures in the high resolution image instead of the whole image. They aim at selecting a specific structure of interest in the high resolution image and allow the system to retrieve similar structures along with the diagnosis information for that specific case.

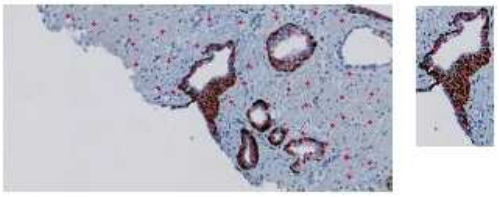

Fig. 1. Structure selection by the pathologist and the resulting image from the database.

In histopathology images, the diagnosis is done by examining a combination of tissue staining, architecture and morphological features. The pathologists look for sets of interesting structures within the images and make the diagnosis by aggregating the information and visual cues for the entire image.

In this paper, we present a sub-image retrieval CBIR system (sCBIR) for high resolution pathology images. The tissues are automatically localized and the background is discarded. Segmentation is not necessary, however, it helps curtail search space considerably. We then perform feature extraction of each image and index them along with diagnosis information in the database. The pathologist is given tools for selection of regions of interest (usually specific structures such as the glands) and perform a query that will retrieve all similar structures along with diagnosis information that will aid the pathologist in diagnosis of this new case.

We utilize a set of scale-invariant features (SIFT) presented by Lowe [6] to identify the structure of interest within the indexed IHC sub-images from the database that are most similar to the selected sub-image by the pathologist.

The remainder of this paper is organized as follows: Section 2 covers the related work and section 3 illustrates our data, sCBIR system generation, and indexing. Our query algorithm is described in section 4. Experimental results are shown in section 5 and we conclude in section 6.

This work was supported in part by the New York State Foundation for Science, Technology and Innovation (NYSTAR) and BioImagene Inc.

2. RELATED WORK

Many researchers have been trying to build CBIR systems with various types of similarity metrics and levels of image indexing. Shyu et al. [2] presented a comparative validation on localized versus global features for CBIR on Computed Tomography (CT) images. They require manual delineation of the pathology bearing regions (PBR) of the images for preparation of the database which adds large burden on radiologists.

Pass et al. [7] presented a similarity metric based on Global Color Histograms (GCH) to classify each pixel in a color bucket (region) as either coherent or incoherent based on its similarity to a large similarly-colored region. They validated their metric on a CBIR system for natural color images which have small sizes compared to pathology images. Wang et al. [8] presented a CBIR system based on semantics classification methods, a wavelet-based approach for feature extraction, and integrated region matching based upon image segmentation. They applied their CBIR on full pathology image retrieval. However, their indexed images are smaller (cropped) images of the original high resolution image. They soften the matching by allowing one region of an image to be matched to several regions of another image. Zheng et al. [9] presented a CBIR system employing a client/server architecture and utilized four image feature types: color histogram, image texture, Fourier coefficients, and wavelet coefficients, using the vector dot product as a distance metric for similarity measurement. Comaniciu et al. [10] presented a CBIR system for clinical application on whole image similarity by retrieving similar cases and not based on sub images as in our case. They utilize Fourier descriptors for shape features, area, and multiresolution models for texture features. Wang [3] used simple features for building a CBIR system but with more relevant region matching similarity metric (Integrated Region Matching). Few authors investigated sub-image retrieval [4, 5], however, their clinical relevance is different from our problem of interest. For example, Luo et al. [4] presented a sub-image retrieval system for natural color images to retrieve similar images with region of interest query. They presented overlapping blocks for feature extraction that overcomes the possible unwanted segmentation of the target image. However, they based their system only on color features.

In digital pathology, the relevance of CAD systems require sub-image retrieval rather than whole image retrieval. Pathologists are interested in specific structures in addition to the whole image for diagnosis. We propose a sub-image CBIR (sCBIR) system for extraction of structures of interest, from prostate (IHC stained) high resolution images, which are responsible for pathology determination in prostate images. These structures include: single lumen glands, multi-lumen glands, PIN (Prostatic Intraepithelial Neoplasia), blood vessels, and lymphocytes. The glands are the main structure that specifies pathology condition, such as malignant and benign glands as shown in Fig. 2.

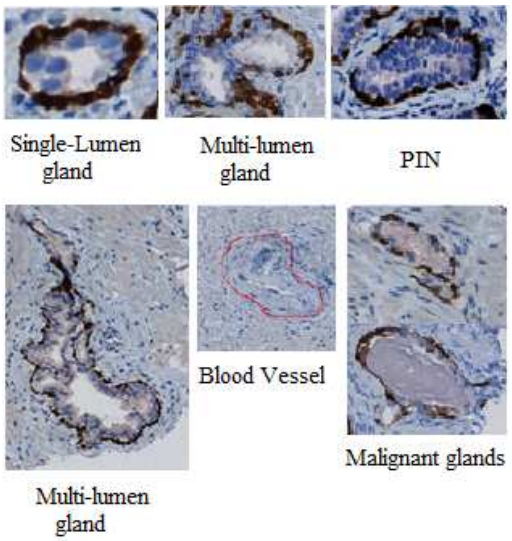

Fig. 2. Various structures of interest in prostate H&E images.

3. DATASET PREPARATION AND INDEXING

Our current dataset consists of 50 IHC stained high resolution pathology images obtained from BioImagene Inc[1]. These images are stored in JPEG2000 format and contain 8 levels of resolution (L_i where $1 \leq i \leq 8$ and L_8 is the highest resolution image),

Because of the comparatively large sizes of JP2 images (around 0.5 GB per case [1]) and the huge space requirement for databases, we do not index the whole JP2 images. Instead, we index the second highest L_7 resolution level of the eight zooming levels that exist in each JP2 image.

We then perform automatic indexing and feature extraction into our database. We input the JP2 images one-by-one to our system as shown in the flow chart in Fig. 3 for indexing. We use the L_7 layer image from the JP2 image for localization of the tissue and thus reducing the unnecessary background of the slide image.

For localization of the tissue, we binarize the image using Otsu threshold [11] and then use the flood fill operation [12] to fill up the holes inside the tissue. This method has been tested only for prostate images. Other pathologies may require different modes of segmentation. However, if there is sufficient contrast between the background and the tissue this procedure is likely to work well.

We next perform the point of interest detection and the SIFT feature extraction as shown in Fig. 3. Each indexed image is then saved into the database along with its SIFT features making the sCBIR ready for new queries.

[1]BioImagene Inc. www.bioimagene.com, 919 Hermosa Court, Sunnyvale, CA 94085

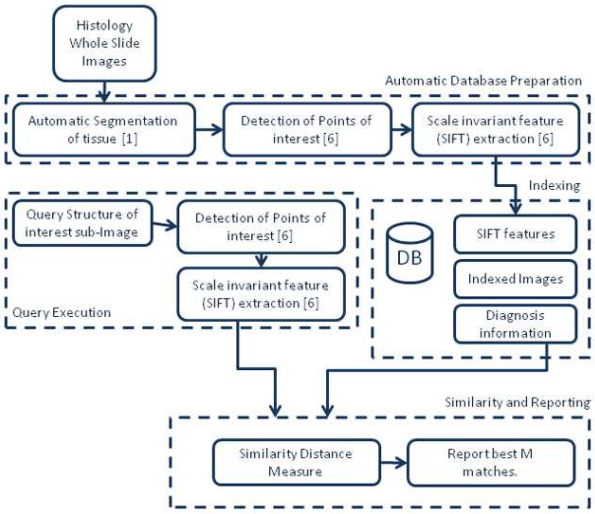

Fig. 3. Flow chart of our sCBIR.

4. FEATURE EXTRACTION AND QUERY ALGORITHM

4.1. Feature Extraction

When the pathologist selects a sub-image $\mathtt{Q}$ that contains a structure of interest, we execute point detection as shown in Fig. 3 to obtain the points $\mathtt{P}_N$ of size N which depends on the selected structure. We use Lowe's [6] method for point detection where the scale-invariant features (SIFT) are efficiently identified by using a staged filtering approach. The first stage identifies key locations in scale space by looking for locations that are maxima or minima of a difference-of-Gaussian $\mathtt{G}(x, y, \sigma)$ function convolved with the image $\mathtt{I}$:

$$\mathtt{D}(x, y, \sigma) = [\mathtt{G}(x, y, k\sigma) - \mathtt{G}(x, y, \sigma)] * \mathtt{I}(x, y) \quad (1)$$

where $\mathtt{D}(x, y, \sigma)$ is the difference function, $\mathtt{G}(x, y, k\sigma)$ and $\mathtt{G}(x, y, \sigma)$ are Gaussian functions with different sigma (k is constant), σ is the Gaussian parameters, $\mathtt{I}(x, y)$ is the image, and x, y are coordinates in the image $\mathtt{I}$.

Each point $\mathtt{P}$ is used to generate a SIFT feature vector $\mathtt{S}_\mathtt{P}$ that describes the local image region sampled relative to its scale-space coordinate frame. The features achieve partial invariance to local variations, such as affine or 3D projections, by blurring image gradient locations. This approach is based on a model of the behavior of complex cells in the cerebral cortex of mammalian vision. SIFT finally gives us a 128 element feature vector for each key point.

4.2. sCBIR Query Algorithm

Our system allows the pathologist to manually select a region of interest image (query image) $\mathtt{Q}$. Then the system computes the set of points of interest $\mathtt{P}_\mathtt{Q}$. We then compute the set of scale-invariant feature vector SIFT [6] for each point $\mathtt{P}_{\mathtt{Q}_i}$. Fig. 1 shows a sample query (right) and an indexed image (left) from the database that contains the query image. The red dots in the image are the key points. Our algorithm is described below as follows:

1. We run a nearest neighbor search for each point $\mathtt{P}_{\mathtt{Q}_i}$ in the query image $\mathtt{Q}$ against all points $\mathtt{P}_\mathtt{I}$ in the database and maintain the closest j points giving a single row i in matrix D:
$$\mathtt{D}_i = \{\min_{\mathtt{P}_\mathtt{I}} |\mathtt{P}_{\mathtt{Q}_i} - \mathtt{P}_\mathtt{I}|\}_j \quad (2)$$
2. Each row in $\mathtt{D}$ represents the closest j points from the database. We then center a window $\mathtt{W}_k$ that is twice the size of the query image $\mathtt{Q}$ in the image that containts this point j. We exclude any point in D that lies inside the window $\mathtt{W}_k$ from any further calculations.
3. Calculate the score value $\mathtt{O}_{\mathtt{W}_k}$ of the sub-image $\mathtt{W}_k$ by counting the number of points in this window. We select the sub-image window $\mathtt{W}_k$ with lower average eucledian distance to the query window for windows with same score values.
4. We repeat the previous step producing a set of sub-images $\mathtt{W} : \{\mathtt{W}_k : 1 \leq k \leq \mathtt{K}\}$ where $\mathtt{K}$ is the final number of sub-images produced. Each sub-image $\mathtt{W}_k$ is stored along with its score value $\mathtt{O}_{\mathtt{W}_k}$ and its average eucledian distance. We omit windows where $\mathtt{O}_{\mathtt{W}_k}$ is less than 10% of $\mathtt{P}_\mathtt{Q}$ to curtail the size of the result.
5. The system will retrieve an ordered set $\mathtt{M}$ of sub-images $\mathtt{W}$ which has the least score values $\mathtt{O}_\mathtt{W}$.

5. EXPERIMENTAL RESULTS

We present results of our sCBIR on ten prostate pathology cases by building a full database and running our sCBIR building procedure shown in Fig. 3. The original high resolution images requires about 0.5 GB per case because they contain eight levels of resolution. The second highest level $\mathtt{L}_7$ which we use for localization and sub-image matching has an approximate average resolution of 7500 x 5000 pixels. After performing the localization step on $\mathtt{L}_7$, we index and use those images for feature extraction and database preparation.

To evaluate the effectiveness of our sCBIR we manually marked similar areas within an image for every query. These regions were not ranked in any particular order. We considered a query result of our sCBIR to be a hit if it overlaped with what we marked and a miss if it did not. After we prepare our index, we apply query images containing specific structures that we mention in Fig. 2. We evaluate our system by comparing the query results of our sCBIR system to the manually generated query results. Results for two such queries are shown in Fig. 4. Fig. 4(a) shows a query for a PIN and its top 3 results. The first retrieved image (second

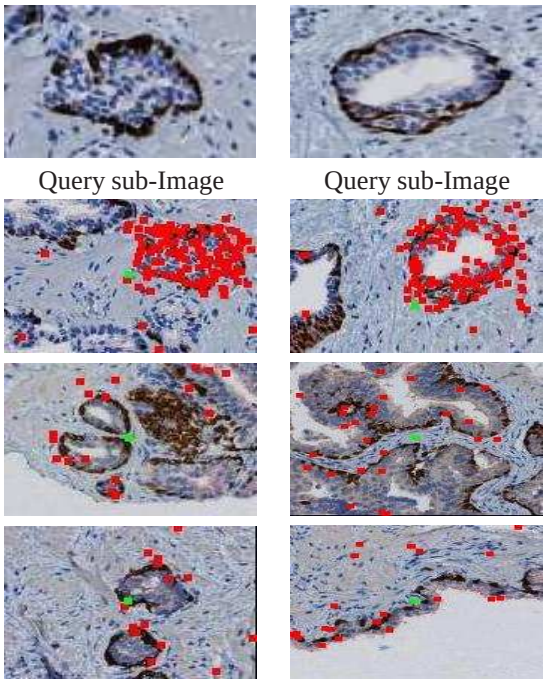

Fig. 4. Sample queries for (a) PIN and (b) single-lumen gland.

image) is the top ranked image which is the query image because the image itself is in the database. The second query example shown in Fig. 4(b) queries a single-lumen structure with the first hit being same query image as it is in the database. The green dot represents the center of the sub image and the red dots represent all the points contained in that sub image.

Table 1. Hits/Misses for 3 queries.

Results	Total Hits	Total Miss	Accuracy
Top 1	3/3	0/3	100%
Top 5	12/15	3/15	80%
Top 10	21/30	9/30	70%
Top 15	29/45	16/45	64.4%

From table 1 it is clear that our system correctly matches the query sub-image to be the given sub-image in all cases. As we move to "Top x" matches, the accuracy of the system goes down with increasing "x". A possible reason for this is that the manually marked similar reqions are not very accurate as well. Clearly we need to have the "top x" results to be validated by at least two pathologists to increase the confidence of the manual results.

6. CONCLUSION

We proposed a Content based sub-Image retrieval System (sCBIR) for clinical pathology computer aided diagnosis system. We automatically prepare a database of pathology images which are JPEG2000 (JP2) compressed images that contain eight layers of resolutions. We performed automatic segmentation and automatically extracted the scale-invariant features from each image and index it into our database. We presented our search algorithm for a given query sub-image and the retrieval of the best ranked sub-images.

Our sCBIR system relies on the clinical practicality of CAD systems for pathology by allowing the pathologist to select a region of interest which is usually a structure that has clinically significant role in pathology diagnosis. Our sCBIR retrieves the top ranked sub-images along with the stored diagnosis (if any) for that sub-image. We validated our sCBIR by querying on structures of interest and comparing results with manually selected matches.

7. REFERENCES

[1] Raja' S. Alomari, Ron Allen, Bikash Sabata, and Vipin Chaudhary, "Localization of tissues in high resolution digital anatomic pathology images," in *In Proc. of the SPIE medical imaging 2009*, Feb 2009.

[2] Shyu Brodley Kak, C. R. Shyu, C. E. Brodley, A. Kosaka A. C. Kak, A. Aisen, and L. Broderick, "Local versus global features for content-based image retrieval, content-based access of image and video libraries," in *Proc. of IEEE Workshop on of Content-Based Access of Image and Video Databases*, 1999, pp. 30–34.

[3] James Z. Wang and M.S. Math, "Pathfinder: Multiresolution region-based searching of pathology images using irm," in *In Proc. Of AMIA Symp.*, 2000, pp. 883–887.

[4] Jie Luo and Mario A. Nascimento, "A content based sub-image retrieval via hierarchical tree matching," in *In Proc. of MMDB04*, 2004.

[5] M. Coutaud, P. Bonnet, A. Joly, R. Enficiaud, N. Boujemaa, and D. Barthlmy, "Advances in taxonomic identification by image recognition with the generic content-based image retrieval ikona," in *In Proc. of Conf. on Biodiversity Informatics*, 2009, To appear.

[6] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, , no. 2, pp. 91–110, 2004.

[7] Greg Pass, Ramin Zabih, and Justin Miller, "Comparing images using color coherence vectors," in *In Proc. of the 4th ACM conf. on Multimedia*, 1999, pp. 65–73.

[8] James Z. Wang, Jia Li, and Gio Wiederhold, "Simplicity: Semantics-sensitive integrated matching for picture libraries," *IEEE Trans. On Pattern Analysis and machine intelligence*, vol. 23, no. 9, 2001.

[9] Lei Zheng, Arthur W. Wetzel, John Gilbertson, and Michael J. Becich, "Design and analysis of a content-based pathology image retrieval system," *IEEE Trans. On Inf. Tech. in Biomedicine*, , no. 4, 7 2003.

[10] Dorin Comaniciu, David Foran, Peter Meer, and Peter Meer, "Image-guided decision support system for pathology," *Machine Vision and Applications*, 1999.

[11] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. On Systems, Man and Cybernetics*, vol. 9, pp. 62–66, 1979.

[12] P. Soille, "Morphological image analysis: Principles and applications," 1999, pp. 173–174.