Region-based Retrieval of Biomedical Images*

James Z. Wang†
Biomedical Informatics and Computer Science
Stanford University
wangz@cs.stanford.edu

ABSTRACT

Searching digital biomedical images is a challenging problem. Prevalent retrieval techniques involve human-supplied text annotations to describe image contents. Biomedical images, such as pathology slides, usually have higher resolution than general-purpose pictures, making it additionally difficult to index. Precise object segmentation is also extremely difficult and is still an open problem. We are developing a multiresolution region-based retrieval system for high-resolution biomedical image databases. The system, based on wavelets, the IRM (Integrated Region Matching) distance, and image classification, is highly robust to inaccurate image segmentation and various visual alterations. Tested on a database of more than 70,000 pathology image fragments, the system has demonstrated high accuracy and fast speed.

1. INTRODUCTION

In biomedicine, content-based image retrieval is critically important in patient digital libraries, clinical diagnosis, clinical trials, searching of 2-D electrophoresis gels, and pathology slides. Most existing content-based image retrieval systems [2, 3, 5, 6, 7, 9] are designed for general-purpose picture libraries such as photos and graphs. In this doctoral dissertation, we present a retrieval system for high-resolution biomedical image databases [10] using wavelet-based [1] multiscale feature extraction and the Integrated Region Matching (IRM) distance [4].

2. THE SYSTEM

2.1 Semantics-sensitive Retrieval

The capability of existing CBIR systems is essentially limited by the way they function, i.e., they rely on only primitive features of the image. Moreover, the same low-level image features and image similarity measures are typically used for images of all semantic types. However, different image features are sensitive to different semantic types. For example, an OCR (optical character recognition) method may be good for graphs commonly found in biomedical educational materials while a region-based indexing approach is much better for pathology and radiology images.

We propose a *semantics-sensitive* approach to the problem of searching image databases. Semantic classification methods are used to categorize images so that semantically-adaptive searching methods applicable to each category can be applied. At the same time, the system can narrow down the searching range to a subset of the original database to facilitate fast retrieval. A biomedical image database may be categorized into "X-ray", "MRI", "pathology", "graphs", "micro-arrays", etc. We then apply a suitable feature extraction method and a corresponding matching metric to each of the semantic classes.

2.2 Feature Indexing

The purpose of content-based indexing and searching for biomedical image databases is very different from that for picture libraries. Users of a general-purpose picture library are typically interested in images with similar object and color configurations at a global scale, while users of a biomedical image database are often interested in images with similar objects at the finest scale.

The feature extraction in our system is performed on multiresolution image blocks (or fragments) of the original images, rather than the original images. In fact, we first partition the images and lower resolution versions of the same image into overlapping blocks. A user may submit an image patch or a sketch of a desired object to form a query. The system attempts to find image fragments within the database to match with the object specified by the user.

A portion of a high-resolution biomedical image is represented by a set of regions, roughly corresponding to objects, which are characterized by color, wavelet-based features, shape, and location. We used the k-means statistical clustering algorithm to obtain fast segmentation. A measure for the overall similarity between images is developed by a region-matching scheme that integrates properties of all the regions in the images. The advantage of using such a "soft matching" is that it makes the metric robust to inaccurate segmentation, an important property that previous work has not solved.

*Project URL: `http://WWW-DB.Stanford.EDU/IMAGE/`

†Principal Advisor: Gio Wiederhold

2.3 Matching

Based on the region representations of the image fragments, the Integrated Region Matching (IRM) distance [4] is used to compute the distance between two image fragments. That is, IRM is a similarity measure between images fragments based on region representations. It incorporates the properties of all the segmented regions so that information about a fragment can be fully used. Region-based matching is a difficult problem because of the problems of inaccurate segmentation. Semantically-precise image segmentation is an extremely difficult process and is still an open problem.

Traditionally, region-based matching is performed on individual regions [5]. The IRM metric we have developed has the following major advantages:

1. Compared with retrieval based on individual regions, the overall "soft similarity" approach in IRM reduces the influence of inaccurate segmentation.
2. In many cases, knowing that one object usually appears with another object helps to clarify the semantics of a particular region.
3. By defining an overall image-to-image similarity measure, the SIMPLIcity system provides users with a *simple* querying interface.

Mathematically, defining a similarity measure is equivalent to defining a distance between sets of points in a high-dimensional space, i.e., the feature space. Every point in the space corresponds to the feature vector, or the descriptor, of a region. Although distance between two points in feature space can be easily defined by various measures such as the Euclidean distance, it is not obvious how to define a distance between two sets of feature points. The distance must correspond to a person's concept of semantic "closeness" of two images.

We argue that a similarity measure based on region segmentation of images can be tolerant of inaccurate image segmentation if it takes all the regions in an image into consideration. To define the similarity measure, we first attempt to match regions in two images. Being aware that segmentation process cannot be perfect, we "soften" the matching by allowing one region of an image to be matched to several regions of another image. Here, a region-to-region *match* is obtained when the regions are significantly similar to each other in terms of the features extracted.

The principle of matching is that the most similar region pair is matched first. We call this matching scheme *integrated region matching* (IRM) to stress the incorporation of regions in the retrieval process. After regions are matched, the similarity measure is computed as a weighted sum of the similarity between region pairs, with weights determined by the matching scheme.

3. EXPERIMENTS

We are developing an experimental image retrieval system, SIMPLIcity (Semantics-sensitive Integrated Matching for Picture LIbraries) [8], to validate these methods on both biomedical and general-purpose image databases. We show that our methods perform much better and much faster than existing methods such as the EMD-based color histogram matching [6] and the WBIIS system based on the Daubechies' wavelets [9]. The system is exceptionally robust to image alterations such as intensity variation, sharpness variation, intentional distortions, cropping, shifting, and rotation. These features are important to biomedical image databases because visual features in the query image are not exactly the same as the visual features in the images in the database. The system has a friendly user interface which is capable of processing a query based on an outside image or a hand-drawn sketch in real-time.

4. ACKNOWLEDGMENTS

I would like to thank my principal advisor, Professor Gio Wiederhold, and my committee members Professors Russ B. Altman, Hector Garcia-Molina, Mu-Tao Wang, and Stephen T.C. Wong, for their guidances on this work. This work was supported in part by the National Science Foundation Grant No. IIS-9817511. I would like to thank the help of Oscar Firschein, Jia Li, Donald Regula and anonymous reviewers.

5. REFERENCES

[1] I.Daubechies, *Ten Lectures on Wavelets*, Capital City Press, 1992.

[2] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, P. Yanker, "Query by image and video content: the QBIC system," *Computer*, vol. 28, no. 9, pp. 23-32, Sept. 1995.

[3] A. Gupta, R. Jain, "Visual information retrieval," *Comm. Assoc. Comp. Mach.*, vol. 40, no. 5, pp. 70-79, May 1997.

[4] J. Li, J. Z. Wang, G. Wiederhold, "IRM: Integrated Region Matching for Image Retrieval," *Proc. of the 2000 ACM Multimedia Conf.*, Los Angeles, October, 2000.

[5] W. Y. Ma, B. Manjunath, "NaTra: A toolbox for navigating large image databases," *Proc. IEEE Int. Conf. Image Processing*, pp. 568-71, 1997.

[6] Y. Rubner, *Perceptual Metrics for Image Database Navigation*, Ph.D. Dissertation, Computer Science Department, Stanford University, May 1999.

[7] J. R. Smith, S.-F. Chang, "An image and video search engine for the World-Wide Web," *Storage and Retrieval for Image and Video Databases V (Sethi, I K and Jain, R C, eds), Proc SPIE 3022*, pp. 84-95, 1997.

[8] J. Z. Wang, J. Li, D. Chan, G. Wiederhold, "Semantics-sensitive Retrieval for Digital Picture Libraries," *D-LIB Magazine*, 5(11), November 1999. http://www.dlib.org

[9] J. Z. Wang, G. Wiederhold, O. Firschein, X.-W. Sha, "Content-based Image Indexing and Searching Using Daubechies' Wavelets," *Int'l J. of Digital Libraries (IJODL)*, 1(4):311-328, Springer-Verlag, 1998.

[10] S.T.C. Wong (ed.), *Medical Image Databases*, Kluwer Academic, Boston, 1998.