Exploring the Use of Image Text for Biomedical Literature Retrieval

Songhua Xu[1], James McCusker[2], Michael Krauthammer[2]

[1]Department of Computer Science, Yale University, New Haven, CT
[2]Department of Pathology & Yale Center for Medical Informatics, New Haven, CT

Abstract

In biomedical publications, figures and images concisely summarize a paper's experimental findings and results. Recent studies have therefore explored the use of images to assist in information retrieval (IR) in biomedicine, mostly based on mining the image caption content. We extend this approach by mining the image text, which refers to the text inside biomedical figures and images. In this work, we discuss the distinct advantages of using image text for biomedical IR and present a prototype search engine implementing the idea.

Introduction

Images contain a paper's key findings. Prior studies have therefore proposed to use image information to assist in biomedical IR (see for example [1]). The overarching idea is to facilitate the retrieval of biomedical articles by making the image content accessible to IR systems. Prior approaches are based on mining the image caption content, or extracting basic image features using image analysis techniques (such as gray level histograms). The image caption text and the image features are then incorporated in the retrieval and classification of biomedical information. However, there exists no IR approach that can retrieve biomedical information by accessing the text within biomedical images. This offers several advantages over searching over image captions alone. First, captions may not contain all the textual information that is contained in the images. Second, image texts are usually very specific, allowing for precise matching of images with related images. We argue that the use of image text (which refers to the text *inside* biomedical images) offers additional advantages for biomedical IR, and present a system for accessing the biomedical literature via image text.

Our Approach and Prototype System

We implemented a prototype system for image and literature retrieval based on image text. We extract image text through image segmentation and Optical Character Recognition (OCR) in biomedical images. Our system has indexed over 140,000 images from public-access biomedical journal papers. A user can compose an image query by specifying the word(s) he expects to appear inside an image, and -optionally- in the image caption, or in the associate paper title and abstract. Once the query is submitted, he will be presented with a page of thumbnails of the images found by our engine. After he clicks on an image, he will be directed to a page pointing to the source paper containing the image. This way, a user can find a paper of interest via the paper's images and figures. The system also provides the user with thumbnails of related images from other papers. After clicking on one of those thumbnails, the user will be redirected to a page depicting the images and figures of the related paper. This way, users can conveniently and intuitively navigate through related papers of interest. User response so far has been overwhelmingly positive. We have investigated several aspects of our system, including the image text extraction performance, and several mechanisms to aid in the image text extraction process. Also, we are performing evaluation studies to measure the benefits of using image text for biomedical information retrieval. Our first results indicate that our approach offers distinct advantages over traditional ways of accessing biomedical images.
Our search engine can be accessed at http://krauthammerlab.med.yale.edu/imagefinder

Conclusion

We propose to use image text for retrieving biomedical images and their associated papers, and present a prototype search engine implementing the idea. System evaluation, and evaluation of image retrieval performance, indicate that our approach is working well, and is capable of retrieving image information that is beyond the reach of traditional IR systems.

References

1. Hearst M, Divoli A, Guturu H. et al. BioText Search Engine: beyond abstract search, Bioinformatics, 2007; 23(16):2196-2197.