Improving OCR Performance in Biomedical Literature Retrieval through Preprocessing and Postprocessing

Songhua Xu [1], **James McCusker** [2], **Martin Schultz** [1], **and Michael Krauthammer** [2]
[1]Department of Computer Science, Yale University
51 Prospect Street, New Haven, CT 06520-8285, USA
[2]Department of Pathology & Yale Center for Medical Informatics
300 Cedar Street, New Haven, CT 06510, USA
{songhua.xu, james.mccusker, martin.schultz, michael.krauthammer}@yale.edu

Abstract

Today's information retrieval (IR) techniques are mostly text-based. As a consequence, some types of information are beyond the reach of text-based IR systems, which fail in situations where textual information can not be easily accessed, e.g. textual information in biomedical images and figures. To tackle such situations, we propose to augment IR systems with the ability to perform optical character recognition (OCR). A principal obstacle is the accuracy of the OCR procedure, which is often error-prone. In our work, we introduce some preprocessing and postprocessing techniques for improving the OCR performance. Our preprocessing stage is concerned with separating texts from graphical elements in an image so that the graphics in the image would not affect the performance of OCR, as today's OCR engines are optimized for dealing with documents without graphical elements. Our postprocessing stage is concerned with a context-based OCR result correction. Experimental results show that these preprocessing and postprocessing techniques can consistently improve the performance of biomedical image OCR in terms of either precision or recall.

1 Introduction

In biomedical publications, figures and images often 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. To study the potential of using image text for information retrieval over the biomedical literature, we developed a prototype search engine based on image text search called *Yale Image Finder*, which is publicly available at (*http://kauthammerlab.med.yale.edu/imagefinder*). In a high-level evaluation of image search performance, we demonstrated that the search engine is capable of retrieving a higher number of relevant images compared to querying against the image caption alone (Xu et al., 2008).

An obstacle to the development of a text-based search engine is the accuracy of the OCR procedure, which is often error-prone. In our work, we introduce some preprocessing and postprocessing techniques for improving the OCR performance. Our preprocessing step involves layout analysis to detect and extract text from surrounding graphical elements. As a result, graphical elements do not degrade the performance of the OCR engine, which is optimized for dealing with documents without graphical elements. Our postprocessing step is concerned with performing a context-based OCR result correction. The key idea is to capture the textual context for each biomedical image. We assume that texts within biomedical images are discussed in their textual context, i.e. in the image caption, in the paragraph that discusses the image, or in the paper that features the image. We thus correct the raw image OCR result by matching it to the terms found in its context. Experimental results show that these preprocessing and postprocessing techniques

consistently improve the performance of biomedical image OCR in terms of either precision or recall.

2 A Prototype Biomedical Literature Search Engine Based on Image Text

Prior studies have proposed to use image information, mostly image caption, to assist in biomedical IR (see for example (Hearst et al., 2007)). We extend this idea and propose to facilitate the retrieval of biomedical articles by making the image content accessible to IR systems. 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 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. For OCR, we used the Image Analysis toolbox (Document Imaging) that is part of Microsoft Office 2003 Professional. Our system has indexed over 100,000 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 is presented with images that are relevant to his query (see *http://krauthammerlab.med.yale.edu/imagefinder*).

We have investigated several aspects of our system, including the image text extraction performance (Xu et al., 2008). Our results indicate that on average, only about 30% of image text is contained in the caption of images, and that for queries that contained two ore more search strings, we were able to retrieve 30% to 175% more images compared to searching over caption alone.

3 Preprocessing and Postprocessing Techniques for Improving OCR Performance

Since our new biomedical literature search engine functions through searching image texts, the OCR performance will critically affect the performance of our search engine. Therefore, we introduce a set of preprocessing and postprocessing techniques for improving OCR performance.

The key idea behind our preprocessing step is to provide customized layout analysis over images published in academic journals, using histogram-based image processing techniques (Lienhart and Wernicke, 2002; Wu et al., 1999). The analysis identifies image text elements, and subjects them to OCR. The text extraction is repeated after turning an image 90 degrees, to allow for the capture of vertical image labels.

The key operation in our postprocessing step is to cross-check extracted image text against the context of the images, and to retain image text which is mentioned in its context. Such context-based correction can effectively minimize false positive results, as intensively discussed in prior studies (Kukich, 1992; Ringlstetter et al., 2007). In our current implementation, we work with two types of image context: one is constituted by all the words from the article that features the image, and the other is constituted by the words in the public accessible articles from PubMed Central. We call image text correction based on the former context "article-based correction", and image text correction based on the latter context "corpus-based correction".

In this study, we evaluate these preprocessing and postprocessing steps, either alone or in combination. The goal is to determine the optimal processing pipeline to extract text from biomedical images. We evaluate the following processing options:

Plain-uncorrected option This option uses raw OCR output without any preprocessing or postprocessing.

Plain-corrected option This option uses article-based correction in the postprocessing stage.

Layout-uncorrected option This option uses layout analysis in the preprocessing stage.

Layout-corrected option This option uses layout analysis in the preprocessing stage and article-based correction in the postprocessing stage.

Corpus-plain-corrected option This option uses corpus-based correction in the postprocessing stage.

Corpus-layout-corrected option This option uses layout analysis in the preprocessing stage and corpus-based correction in the postprocessing stage.

High-recall option This option combines the plain-uncorrected option and layout-uncorrected option.

High-precision option This option combines the results from the plain-corrected option, layout-corrected option, corpus-plain-corrected option, and corpus-layout-corrected option.

The latter two options combine the best preprocessing and postprocessing procedures to either retrieve most of the image text content (high-recall option) or to retrieve image text context with the highest amount of precision (high-precision option).

4 Evaluation

To evaluate the effectiveness of our OCR correction techniques, we conducted two evaluations, where we compared OCR-extracted and corrected image text against manually extracted image text. The first evaluation focused on 343 random images whose captions contain the word "survival"; the other evaluation focused on 362 random images whose captions contain the word "apoptosis". Both evaluations covered typical biomedical images, such as graphs, diagrams and experimental results.

In Figure 1, we report the results for all the pre- and postprocessing correction options, and combinations thereof, as discussed in Section 3. We analyze the performance with respect to different word lengths. One reason for doing so is that in the postprocessing stage, our context-based correction methods are less efficient for shorter words. This can be intuitively understood as short text strings, which have been erroneously extracted from images, are more likely to be coincidentally mentioned in the image context. According to these results, we find that context-based image text postprocessing improves precision significantly. We also observe that layout-analysis based preprocessing improves recall, specifically when combined with plain (raw) OCR processing. This can be seen in our high-precision option, where we pool the results of layout-analysis based preprocessing with plain (raw) processing, and apply various context-based post-processing steps. Using this option, we achieve the best overall performance in terms of F-rate. Our high-recall option offers the best performance for retrieving terms that are actually mentioned in biomedical images.

5 Conclusion

In this paper, we introduce preprocessing and post-processing techniques for improving OCR-based image text extraction. We show that a combination of image layout analysis and context-based image text correction is most beneficial for boosting OCR performance over biomedical images.

Acknowledgement

This research has been funded by NLM grant 5K22LM009255. We thank Nam Tran, ThaiBinh Luong, Sebastian Szpakowski, and Pavithra Shivakumar for manually labeling the image text in the "survival" and "apoptosis" image sets.

References

Marti A. Hearst, Anna Divoli, Harendra Guturu, Alex Ksikes, Preslav Nakov, Michael A. Wooldridge, and Jerry Ye. 2007. Biotext search engine: beyond abstract search. *Bioinformatics*, 23(16):2196–2197.

Karen Kukich. 1992. Technique for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–439.

R. Lienhart and A. Wernicke. 2002. Localizing and segmenting text in images and videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(4):256–268, Apr.

Christoph Ringlstetter, Klaus U. Schulz, and Stoyan Mihov. 2007. Adaptive text correction with web-crawled domain-dependent dictionaries. *ACM Transactions on Speech and Language Processing*, 4(4):1–36.

V. Wu, R. Manmatha, and E.M. Riseman. 1999. Textfinder: an automatic system to detect and recognize text in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11):1224–1229, Nov.

S. Xu, J. McCusker, and M. Krauthammer. 2008. Yale image finder (YIF): a new search engine for retrieving biomedical images. *Bioinformatics, Advanced Access*, July.

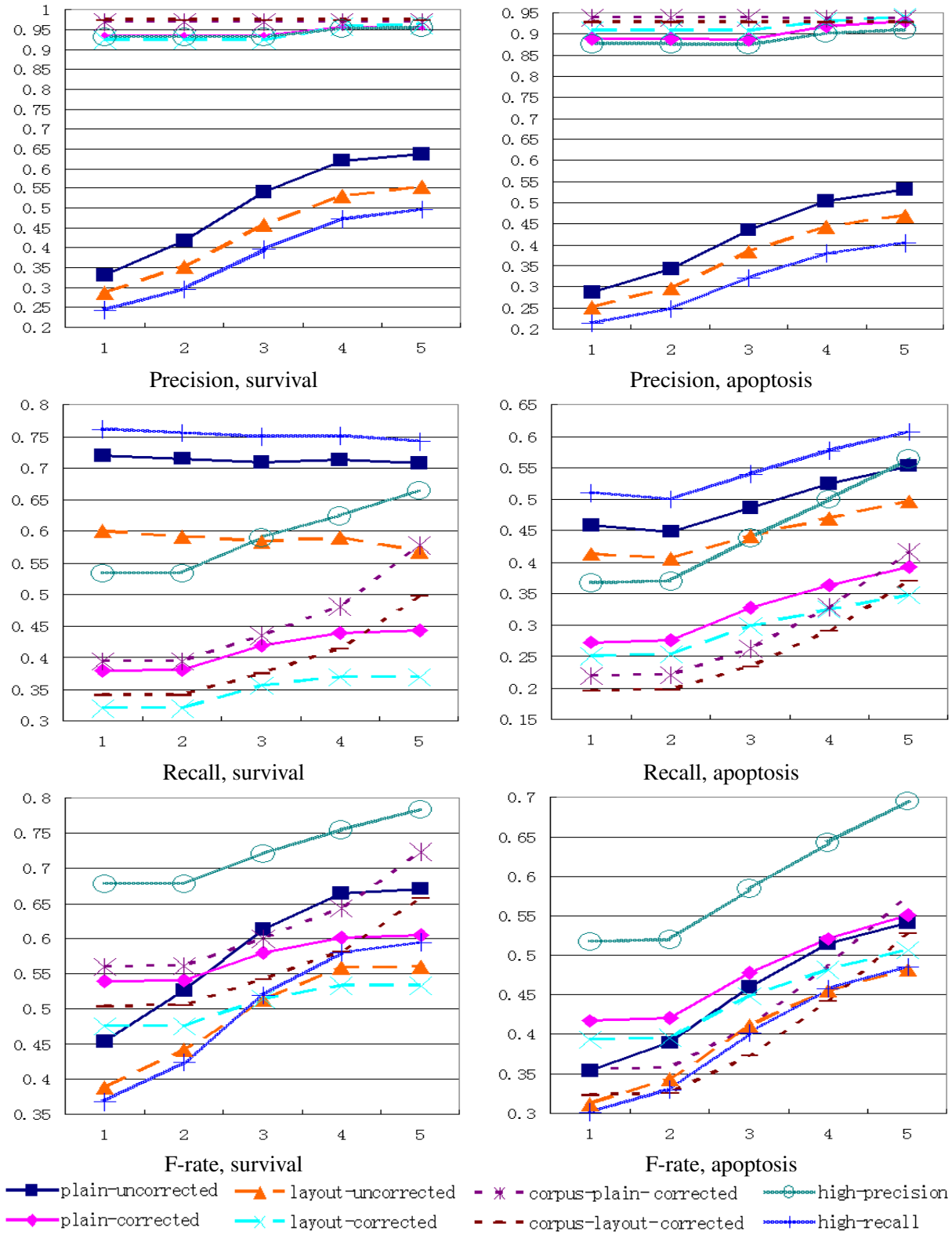

Figure 1: Performance of our method over the survival and apoptosis image sets. Here we show the precision, recall and F-rates (y-axis) for the survival and apoptosis image sets for different pre- and postprocessing methods with respect to different word lengths (x-axis). Results for word length 1 correspond to the overall performance, as we include all words of length 1 and more. From these results, we can see that our high-precision option achieves the best overall performance in terms of F-rate and our high-recall option offers the best performance for retrieving terms that are actually mentioned in biomedical images, i.e. the highest recall.