Boosting Text Extraction From Biomedical Images using Text Region Detection

Songhua Xu and Michael Krauthammer

***Abstract*—In this paper, we show that domain-optimized text detection in biomedical images is important for boosting text extraction recall via off-the-shelf OCR engines. Methodologically, we contrast OCR performance when processing raw biomedical images, compared to preprocessing those images, and performing OCR on detected image text regions only. To quantify OCR extraction results, we rely on a gold standard image text corpus with manually identified image text. To demonstrate the positive effect on biomedical image retrieval, we apply image text detection and extraction to a large corpus of biomedical images in the Yale Image Finder system. We show that improved text extraction results in the retrieval of a larger number of relevant images for a set of domain-relevant keyword searches.**

I. Introduction and Main Idea

In our prior work, we established the critical role of Optical Character Recognition (OCR) for retrieving relevant biomedical images in large image collections [11]. OCR is instrumental in extracting image text (text within images) that is subsequently indexed and made available for image search. An early obstacle of our work was the availability of efficient off-the-shelf OCR engines, which are optimized for extracting text from documents, rather than stand-alone images. We hypothesized that domain-specific preprocessing, i.e. the detection of text regions in biomedical images, and subjecting those regions to the OCR procedure might be beneficial compared to (raw) processing of the whole image. We thus developed a domain-specific text detection program based iterative projection histograms [10].

In this paper, we demonstrate how text detection via iterative projection histograms is helpful for boosting the OCR recall using two off-the-shelf OCR systems, one with intermediate layout capabilities (T-OCR), and one with advanced layout capabilities (M-OCR). To this end, we present an OCR performance evaluation strategy using an image corpus with manually extracted image text. We discuss the qualitative issues involved in creating this corpus, including the problem of low resolution images that are difficult to read by human evaluators (and by OCR engines alike). Another issue is the need for clear guidelines that specify what constitutes an image text string and how to deal with special characters. We further discuss our strategy for determining a positive match between an OCR result and a manually identified image text string. We argue that constrained sloppy matching is better suited to assess the OCR performance, as opposed to an exact string matching procedure.

Using the gold standard image text corpus and the image string matching criteria, we show how our pre-processing (text detection) technique can substantially boost overall T-OCR performance. The results for M-OCR, a more advanced OCR system, are equally promising, but are distinctly different from the T-OCR results. While preprocessing and text detection by itself does not confer improved OCR performance, we find that domain-specific text region detection results in the extraction of additional text, which is not extracted by raw (direct) processing. We thus show that two T-OCR runs (over the raw image, and over detected text regions) and pooling the OCR results from the two runs, results in an improved T-OCR performance, particularly for OCR recall, which is key for improving overall biomedical image retrieval.

II. Related Work

Several teams have proposed tools and methods for retrieving biomedical images via associated image text, such as using abstract sentences [15], or by using advanced NLP over journal abstracts [8]. Our group has advocated the use of image text (text within images) for improving biomedical image retrieval [11], and the use of preprocessing, specifically image text detection, for improving the recall of OCR processing [10]. In the preprocessing step, one is interested in separating image text elements from other elements in images. Hence, our work is closely related to prior work on text region detection in images [9], [3], [1] and videos [2], [4], [5], [6], [7], [14], [16].

III. Improving OCR-Based Image Text Extraction Performance via Domain-specific Text Detection

Since our objective is to improve image retrieval through indexing and searching image text (text within images), the OCR-based image text extraction performance will critically affect the performance of our image search strategy; In this paper, we study the effectiveness of domain-specific text detection for improving OCR performance over biomedical images, and study its effectiveness for overall biomedical image retrieval.

This research has been supported by the National Library of Medicine (grant number 5R01LM009956). Songhua Xu performed this research partly as a Eugene P. Wigner Fellow and staff member at the Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract DE-AC05-00OR22725.

S. Xu is with Oak Ridge National Laboratory, One Bethel Valley Road, Oak Ridge, TN 37831, USA. The work was partly done when he was with the Department of Pathology, Yale University School of Medicine, 300 George Street, New Haven, CT 06511, USA. Email: xus1@ornl.gov.

M. Krauthammer is with the Department of Pathology, Yale University School of Medicine, 300 George Street, New Haven, CT 06511, USA. Email: michael.krauthammer@yale.edu.

978-1-4244-6714-3/10/$26.00 ©2011 IEEE

A. Domain-specific Text Detection

The key idea behind our pre-processing step is to first detect and separate text regions in an image and then individually subject each resulting text region to an OCR process for image text extraction. The image text separately extracted from each text region is pooled, representing the overall image text extraction result for the whole image. In our implementation, we use the iterative and pivoting text region detection algorithm introduced in [10] to identify and separate text regions in biomedical images.

B. OCR Engines

In this study, we used two OCR engines: an OCR engine within the Microsoft Document Imaging package, available in Microsoft Office 2003, and Top OCR, a freely available OCR engine. We selected these two engines after evaluating other public domain OCR engines, including Tesseract, GOCR, and SimpleOCR, for which we were unable to obtain satisfactory image OCR results (data not shown). In the remaining part of this paper, we will refer to the two OCR engine as M-OCR (Microsoft Document Imaging package) and T-OCR (T OCR).

C. Different Text Detection Pre-processing Options

In this study, we evaluate different OCR pre-processing options: a) using text detection before OCR processing, b) raw (direct) OCR processing, and c) combined processing where we pool OCR results from two OCR runs, one with and one without text detection. The goal is to determine the optimal processing configuration to extract text from biomedical images using off-the-shelve OCR engines. We evaluate the following processing options:

- **No text detection** This option uses the raw image as OCR input, without any pre-processing.
- **With text detection** This option uses image text region detection and separation in the pre-processing stage.
- **Combined** This option combines a run with no pre-processing (OCR over raw image) and a run with pre-processing (i.e. text detection before OCR run).

IV. Evaluation Setup

A. Creation of a Gold Standard Image Text Corpus

We were interested in creating a gold standard image text corpus that contained the image text strings from a random collection of biomedical images. After selecting 200 images at random from PubMed Central, we asked several reviewers to manually identify the image text contained in those images. At least two reviewers viewed all images. To achieve consistency across the corpus, we set up guidelines for identifying image text in biomedical images, which are listed in Table I. The guidelines define the nature of an image text string, and what to do about Greek letters and other special characters and strings that consist of only numbers, super-, or subscript characters. Note that this set of image text extraction guidelines is slightly different from the guidelines defined in [10]. The main differences are the treatment of numbers, subscript and superscript letters, which are ignored for the purpose of this study, as these three types of elements are particularly hard to detect using OCR processing, and are rarely queried for in real-live image retrieval scenarios.

TABLE I
GUIDELINES FOR THE MANUAL EXTRACTION OF IMAGE TEXT.

1) The extraction targets are sequences of letters that appear together in an image, such as “bcl-xl(+)”, “abl”, “p<0.01”, and “S.E.RECOVERSOLUTION(i,j,k)”.
2) Except for situations listed below include symbols, and Greek letters. If a symbol cannot be typed, replace it with a space.
3) Ignore words that consist of just numbers.
4) If not attached to a word, ignore operators, such as plus or percentage signs.
5) Do not ignore these operators if attached to a word, such as in “vif(+)”.
6) Always ignore subscript and superscript.
7) Do not ignore forward slashes, and dashes, such as in “MMC-transgenic”.

B. Evaluation Strategy

We also needed to decide on the correct matching procedure between manually recognized image text and those recognized through OCR. An exact matching procedure seemed a straightforward choice for assessing OCR performance. We believed, however, that an exact matching procedure might unfairly penalize the OCR performance in situations where human reviewers separated two image text strings, while the OCR engine would concatenate them together (and vice versa). Therefore, we conceived an alternative matching procedure called constrained sloppy matching. In the latter, we allowed for substring matching between the OCR result and the manually extracted result in situations where at least 50% of the characters overlapped. As can be seen later, the two different evaluation strategies resulted in approximately 5% performance difference, with the sloppy matching reporting slightly better results than the exact matching.

V. Experimental Results

A. Gold Standard Image Text Corpus

We created a gold standard image text corpus as a basis for evaluating image text extraction capabilities of an OCR procedure. After selecting 200 images at random from the PubMed Central image corpus, we encountered some initial obstacles, notably the fact that some of the images could not be read by human reviewers due to the low resolution of the image. Some images were only partially readable. We had two options: retain those images in the corpus with missing human annotation or remove them from the image set. We opted for the latter, assuming that the OCR engine would not recognize any sensible text strings. The decision to remove these images may introduce a slight bias due to the possibility that the OCR engine might have recognized false positive text strings in those low resolution images; as a result, the actual precision numbers might be slightly lower than reported. Our final image text corpus consisted of 161 random images, corresponding to 2445 image text strings, of which 70.84% were not found in the associated image captions.

TABLE II
PERFORMANCE SUMMARY FOR WORDS CONTAINING TWO AND MORE CHARACTERS. THE MAPPING BETWEEN THE AUTOMATED AND MANUAL EXTRACTION RESULTS IS SET TO 50% OVERLAP.

OCR Engine	Pre-processing option	Precision	Recall	F-rate
M-OCR	No text detection	0.404	**0.655**	0.499
	With text detection	0.353	0.550	0.430
	Combined	0.296	**0.707**	0.417
T-OCR	No text detection	0.323	**0.241**	0.276
	With text detection	0.234	**0.310**	0.266
	Combined	0.225	**0.395**	0.287

TABLE III
PERFORMANCE SUMMARY FOR WORDS CONTAINING TWO AND MORE CHARACTERS. THE MAPPING BETWEEN THE AUTOMATED AND MANUAL EXTRACTION RESULTS IS SET TO 100% OVERLAP.

OCR Engine	Pre-processing option	Precision	Recall	F-rate
M-OCR	No text detection	0.382	**0.621**	0.473
	With text detection	0.306	0.482	0.374
	Combined	0.258	**0.667**	0.372
T-OCR	No text detection	0.288	**0.201**	0.236
	With text detection	0.180	**0.239**	0.206
	Combined	0.174	**0.319**	0.225

B. OCR Performance With and Without Image Text Detection

We evaluated the performance of two off-the-shelve OCR engines, which we tested in two different settings. The first setting corresponded to an OCR run with text detection, and the second to one without text detection. We also evaluated a third setting, where we pooled OCR results from two OCR runs, one with and one without image text region detection. Using our gold standard image text corpus, we generated the following statistics: text extraction recall, precision, and F-score.

As discussed above, we call the straight processing of images using the OCR engine, the *No text detection* option, and the processing of images by first segmenting out image text elements, the *With text detection* option. Further, we call the combination of these two settings, one with and one without pre-processing, the *combined* option. As can be seen in Table II, the *With text detection* option performs similar to the direct OCR-processing (*No text detection* option), with a slightly lower F-score of 1-7%. The advantage of the pre-processing routine lies in the increased recall, as can be seen in the 7% increase in recall for T-OCR. While recall was not increased for M-OCR using the *With text detection* option, the *combined* option showed higher recall of 5 %. The *combined* option achieves a 15% increase in recall for T-OCR. In other words, our pre-processing routine identifies image text elements not recognized by the off-the-shelf OCR engines, and combining the results of two OCR runs, one with and one without pre-processing, boosts overall image text recall. We also report evaluation data for the case when the matching criteria is set to exact matching, summarized in Table III, which show similar results as reported when using a relaxed matching criteria.

We are also interested in measuring algorithm performance for different word lengths (data not shown). We observe the following: image text extraction performance is better for longer words. Also, the *combined* option achieves significantly higher image text extraction recall across most image text string lengths. An exception is seen for words of character length 7, for which we see no significant performance gain. Nevertheless, the *combined* option consistently shows higher recall across all word lengths.

Finally, we examined the role of pre-processing (i.e. text detection) on actual image retrieval in our YIF search system [11], which uses the M-OCR engine to extract and index image text across several hundred thousand of images. For a few selected queries, we measured the additional images retrieved and the image retrieval precision when using text detection as a preprocessing step before OCR text extraction. We contrast a situation where we use no pre-processing at all with a situation where we use the combined option of our pre-processing procedure (the combined option results in superior performance for M-OCR). For each search term, we calculate the number of additional images retrieved (Δ) when using the combined option. As can be seen in table Table IV, using the combined pre-processing option, we achieve around a 10 to 14% improvement in terms of relevant images retrieved. This is achieved at 100% image retrieval precision.

VI. CONCLUSION AND DISCUSSIONS

Biomedical image search is increasingly being recognized as a complementary approach for accessing the biomedical literature. As shown in our prior work [13], [12], [11], extracting text embedded in biomedical images is an efficient way to improve image retrieval recall. In this paper, we have shown that the pre-processing of biomedical images is essential for improving the recall of text extraction and image retrieval using off-the-shelf OCR engines. Our pre-processing procedure essentially involves a layout analysis process to detect and separate text regions from surrounding graphical regions. We propose that with such a pre-processing procedure, commercial OCR engines may perform better on images with mixed textual and graphical content.

To explore the effectiveness of our pre-processing technique for image text extraction quantitatively, we created a gold standard image text corpus, with manually extracted image text. We discussed the need for clear guidelines on how to define image text strings, and raised the issue of low-resolution images, which we excluded from the gold standard corpus. This decision, motivated by the fact that manual evaluators cannot extract text from low-resolution images, may introduce a slight bias when measuring OCR extraction precision.

We also needed to determine the best way to compare the automated with the manual text extraction results. A sloppy and an exact matching procedure produced similar performance numbers, with the exact matching resulting in roughly 5% lower F-scores. We introduced the concept of constrained sloppy matching to obtain a more complete understanding of the image text extraction performance using our pre-processing technique. Consider the case where three embedded text strings, 'PMA," "+," and "ionomycin," appear side-by-side inside an image. It is not clear whether the authors of the image put a space between any of the three strings. This may result in a situation where the OCR procedure concatenates

TABLE IV
PERFORMANCE STATISTICS OF IMAGE RETRIEVAL USING THE YIF SYSTEM. WE SHOW THE NUMBERS OF TOTAL AND RELEVANT IMAGES RETRIEVED BY THE YIF SYSTEM FOR THREE TYPICAL QUERIES.

No.	Query	Pre-processing option	Total images	Relevant images	Precision
1	apoptosis	No text detection	355	355	100%
		Combined	390	390	100%
		Δ	35 (9.9%)	35 (9.9%)	
2	insulin	No text detection	297	297	100%
		Combined	328	328	100%
		Δ	31 (10.4%)	31 (10.4%)	
3	vegf	No text detection	231	231	100%
		Combined	263	263	100%
		Δ	32 (13.9%)	32 (13.9%)	

the strings, while the human evaluators separate them (during manual evaluation). As a result, the system might be penalized for not finding the "correct" string using the exact matching criteria. Using sloppy matching, the system is not penalized for separating (or concatenating) strings, as partial overlap is counted as true positive. In this paper, we reported numbers for both measures (exact matching and sloppy matching).

Using the gold standard image text corpus, we compared the automatic image OCR result with the manual image text extraction results. We used two off-the-shelf OCR engines, one with a standard image layout capabilities (T-OCR), and one with an advanced image layout capability (M-OCR). For T-OCR, we showed that pre-processing of biomedical images, that is, detecting and separating image text elements from graphical elements before OCR processing, improves image text extraction recall. For both T-OCR and M-OCR, combining the output of two OCR runs, one with and one without image text region detection and separation, improved OCR recall.

Image text extraction precision is generally lower when using pre-processing. However, the lower precision is mostly due to noisy OCR results, which are counted as false positives. These strings often consist of sequences of random symbols, operators, and other special characters. Users of the YIF system will not search for such strings (i.e., the presence of these strings in images). As a consequence, these false positive results will not affect the overall image search performance. This is shown in Table IV, where we demonstrate that pre-processing in the YIF system results in the retrieval of approximately 10% more images without reducing the retrieval precision.

Based on the results of our study presented in this paper, we conclude that our pre-processing routine is helpful for boosting image text extraction recall when using off-the-shelf OCR engines, and that such a pre-processing technique will not noticeably affect the precision of end-user image retrieval performance.

ACKNOWLEDGEMENT

We thank Martin Schultz and David Gelernter for many stimulating discussions during this research. This work is based on Chapter 4 of S. Xu's Ph.D. thesis—"A New Biomedical Image Search and Visual Literature Navigation System," Department of Computer Science, Yale University, 2010, supervised by Michael Krauthammer, Martin Schultz and David Gelernter.

REFERENCES

[1] E. D. Haritaoglu and I. Haritaoglu. Real time image enhancement and segmentation for sign/text detection. In *Proceedings of International Conference on Image Processing*, volume 3, pages 993–996, 2003.

[2] K. Jung, K. I. Kim, and A. K. Jain. Text information extraction in images and video: a survey. *Pattern Recognition*, 37(5):977–997, 2004.

[3] K. C. Jung, K. I. Kim, T. Kurata, M. Kourogi, and J. H. Han. Text scanner with text detection technology on image sequences. In *Proceedings of the 16th International Conference on Pattern Recognition*, volume 3, pages 473–476, 2002.

[4] D. Kim and K. Sohn. Static text region detection in video sequences using color and orientation consistencies. In *Proceedings of 19th International Conference on Pattern Recognition*, volume 1–6, pages 1754–1757, 2008.

[5] W. Kim and C. Kim. A new approach for overlay text detection and extraction from complex video scene. *IEEE Transactions on Image Processing*, 18(2):401–411, 2009.

[6] M. H. Li and C. H. Wang. An adaptive text detection approach in images and video frames. In *Proceedings of 2008 IEEE International Joint Conference on Neural Networks*, volume 1–8, pages 72–77, 2008.

[7] C. W. Ngo and C. K. Chan. Video text detection and segmentation for optical character recognition. *Multimedia Systems*, 10(3):261–272, 2005.

[8] P. K. Tulipano, Y. Tao, W. S. Millar, P. Zanzonico, K. Kolbert, H. Xu, H. Yu, L. Chen, Y. A. Lussier, and C. Friedman. Natural language processing and visualization in the molecular imaging domain. *Journal of Biomedical Informatics*, 40(3):270–281, 2007.

[9] J. Wu, S. L. Qu, Q. Zhou, and W. Y. Wang. Automatic text detection in complex color image. In *Proceedings of International Conference on Machine Learning and Cybernetics*, volume 1–4, pages 1167–1171, 2002.

[10] S. Xu and M. Krauthammer. A new pivoting and iterative text detection algorithm for biomedical images. *Journal of Biomedical Informatics*, 43(6):924–931, 2010.

[11] S. Xu, J. McCusker, and M. Krauthammer. Yale image finder (YIF): a new search engine for retrieving biomedical images. *Bioinformatics*, 24(17):1968–1970, 2008.

[12] S. Xu, J. McCusker, M. Schultz, and M. Krauthammer. Improving OCR performance in biomedical literature retrieval through preprocessing and postprocessing. In *Proceedings of International Symposium on Semantic Mining in Biomedicine (SMBM '08)*, pages 161–164, 2008.

[13] S. Xu, M. Schultz, and M. Krauthammer. Using textual context for improving OCR performance in biomedical literature retrieval. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology (ISMB '08)*, Toronto, Canada, 2008.

[14] Q. Ye, Q. Huang, W. Gao, and D. Zhao. Fast and robust text detection in images and video frames. *Image and Vision Computing*, 23(6):565–576, 2005.

[15] H. Yu and M. Lee. Accessing bioscience images from abstract sentences. *Bioinformatics*, 22(14):e547–e556, 2006.

[16] C. J. Zhu, Y. X. Ouyang, L. Gao, Z. Y. Chen, and Z. Xiong. An automatic video text detection, localization and extraction approach. *Advanced Internet Based Systems and Applications*, 4879:1–9, 2009.