

# Extraction and Labeling High-resolution Images from PDF Documents

Suchet K. Chachra, Zhiyun Xue, Sameer Antani, Dina Demner-Fushman, George R. Thoma  
Lister Hill National Center for Biomedical Communications  
U. S. National Library of Medicine, Bethesda, MD 20894

## ABSTRACT

Accuracy of content-based image retrieval is affected by image resolution among other factors. Higher resolution images enable extraction of image features that more accurately represent the image content. In order to improve the relevance of search results for our biomedical image search engine, Open-I, we have developed techniques to extract and label high-resolution versions of figures from biomedical articles supplied in the PDF format. Open-I uses the open-access subset of biomedical articles from the PubMed Central repository hosted by the National Library of Medicine. Articles are available in XML and in publisher supplied PDF formats. As these PDF documents contain little or no meta-data to identify the embedded images, the task includes labeling images according to their figure number in the article after they have been successfully extracted. For this purpose we use the labeled small size images provided with the XML web version of the article. This paper describes the image extraction process and two alternative approaches to perform image labeling that measure the similarity between two images based upon the image intensity projection on the coordinate axes and similarity based upon the normalized cross-correlation between the intensities of two images. Using image identification based on image intensity projection, we were able to achieve a precision of 92.84% and a recall of 82.18% in labeling of the extracted images.

**Keywords:** Image matching, PDF image extraction

## 1. INTRODUCTION

Image retrieval quality in content based image retrieval (CBIR) is directly affected by the quality of image features which, in turn, is affected by the image resolution. Using higher resolution images as input provides better image features. Open-I<sup>1</sup> (<http://openi.nlm.nih.gov>) is a multimodal biomedical image retrieval system developed by the National Library of Medicine<sup>®</sup> (NLM<sup>®</sup>). It indexes the open-access subset of biomedical articles available in PubMed Central<sup>®</sup> (PMC) site hosted by NLM. The site hosts over 500,000 articles and 1.6 million images extracted from them. The articles are supplied in XML and PDF formats. Figures from the article are available as JPEG compressed Web-resolution images. These figures need to be processed for image feature extraction. This processing includes splitting multi-panel figures into individual sub-figures; classification of images by type (e.g., X-ray or photography)<sup>2</sup>; and clustering images for faster retrieval. Accuracy in all of these tasks can be improved through use of high-resolution images. Unfortunately, high-resolution images may not always be available from the publisher in a digitally downloadable format. It is vital to consider all sources of better quality, high-resolution images before choosing standard or low-resolution images available in PMC for feature extraction and indexing. Although, most of the Web sites still use standard resolution images and graphics, one area where better quality, high-resolution and content-rich images have been prevalent is the digital print media. Portable Document Format (PDF) is an open standard published by the International Standards Organization (ISO)<sup>3</sup> and has been the de-facto standard for distributing high quality textual content, often containing images and graphics. Given the popularity and wide acceptance of the PDF format, software to extract images and other embedded graphics from PDF documents has been developed for several platforms.

In this paper, we propose the extraction and identification of high-resolution images from the PDF documents bundled with open-access PMC biomedical articles. We also report on how we achieved the labeling of the extracted images using two different approaches. To be clear, for our task it is insufficient to find “visually similar” images, as in traditional CBIR. Rather, we seek exact matches between low and high resolution versions of the same image since they are being used to index the biomedical articles. We expect that the method will benefit the image search in Open-I by providing better quality, and higher resolution images for indexing. Our ongoing research aims to enable cross-modal (text + image) retrieval on sub-region level and we have developed a query expansion framework in image retrieval domain based on local and global analysis<sup>4</sup>. High-resolution images will also be valuable for extracting image features from image sub-regions within the images.

## 2. BACKGROUND

Open-I is a next-generation information retrieval engine, which is unique in its ability to index both text and images in the articles. The system uses the Essie search engine developed at NLM that is also used in ClinicalTrials.gov (<http://clinicaltrials.gov/>). The biomedical articles contained in the Open-I collection are obtained from the open access subset of PMC. Unfortunately, PMC does not provide high quality images in the downloadable article bundles. However, most articles in their bundle are accompanied by a publisher-supplied PDF document of the original biomedical article along with embedded high-resolution versions of article images. In some cases, these high-resolution images were three times as detailed as the web-resolution image files provided in the bundle. The detail in the embedded high-resolution images provided a compelling motivation to develop a method to extract and label the images. In other words, our task was “given a figure obtained from PubMed Central, to identify the corresponding high-resolution figure from the PDF document”. A corresponding figure is one where the figure-content is equivalent while allowing likely intensity, aspect-ratio, and border differences.

There are several commercially and freely available software tools that extract images and graphics from PDF documents. We tried five such tools, both open source and commercial, for extracting images: *pdfimages*, *ICEPDF*, *JPedal*, *PDFBox*, *PDFexpress*, and *Adobe Acrobat*. However, a common and significant problem in using these tools is their inability to extract complete figures from PDF documents. We observed that several elements of the image such as, labels, annotation marks, and other overlaid graphics, were often missing from the extracted images. In addition, the PDF documents contained little or no meta-data that could be used to identify and label the extracted images. Figure 1 shows four example figures (rows (i) to (iv)) as they appear in the article PDF (column ‘a’) and the extracted output using one of the tools (column ‘b’) to illustrate common problem with existing tools while extracting high-resolution images such as missing elements, text or legends, or having other problems which would make image identification and labeling erroneous. In row (i) we see that the output missed item ‘C’ (graph) in the figure and only extracted the images. In row (ii), the PDF image extractor obtained three individual images instead of one multi-panel figure and the text labels were missing. In row (iii), some legends were missing and the image color was reversed. In row (iv) all text was missing.

Extracting images from PDF documents has been a topic of research interest. Images that include complex vector graphic elements, text, and other pictorial graphic elements are particularly challenging. Xu<sup>5</sup> proposed a method to segment graphics embedded in a PDF document using a layer based document analysis method. Shao<sup>6</sup> proposed a method to recognize and classify diagrams in vector-based PDF documents. Lin<sup>7</sup> proposed a method to identify mathematical formulas using rule-based and learning-based methods. The solutions reported in these articles are specific to their problem domain, and software tools and algorithms are either not available or easily replicable. We aim to use off-the-shelf tools to develop reliable algorithms for extraction and labeling of complex graphics from PDF documents.

## 3. METHODS

### 3.1 Overview

In this section we describe our methods for extraction and labeling of high-resolution figures extracted from biomedical articles in PDF. Figure 2 shows the steps involved in the extraction and labeling of high-resolution images from source PDF documents. The source PDF documents often include high-resolution version of the images that are down-scaled to fit the page or column width. Some of these PDF documents also contain additional images such as borders, publisher banners, and empty spacers (blank white graphics) that are of little interest to us. Additionally, as PDF documents can be embedded with images and graphics in different formats, format conversion is performed to convert the extracted images to the JPEG format. Each extracted image is then compared to the labeled low-resolution images provided in the PMC data set as part of the article bundle. The extracted image is labeled when a match is found. We have developed two methods for image matching: one is based on image intensity projections on the coordinate axis, and the other is based on normalized cross-correlation. We discard the remaining images for which no match is found using our method.

(i)		
(ii)		<p style="text-align: center;">(obtained three individual images)</p>
(iii)		
(iv)	<p style="text-align: center;">(a)</p>	<p style="text-align: center;">(b)</p>

Figure 1. Left column (a): Example figure images from open-access biomedical articles in PMC; Right column (b): output of PDF image extraction tool from corresponding PDF documents. Each row exhibits a shortcoming in the tool.

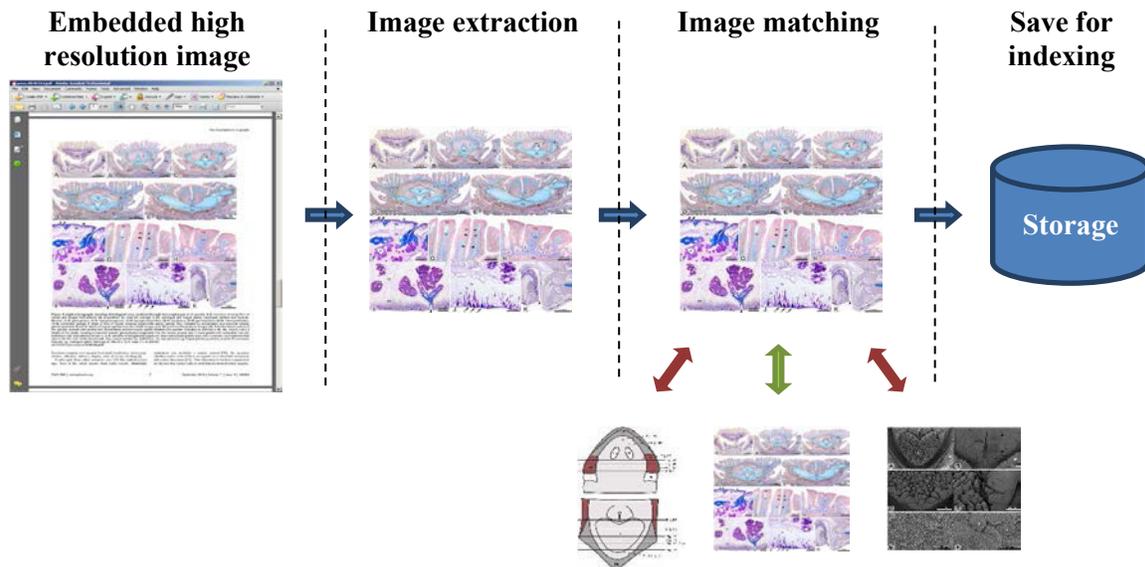


Figure 2. Image extraction and identification process.

### 3.2 Extraction of Images from PDF documents

To perform the extraction of images embedded in the PDF documents, we used an open-source command-line utility called *pdfimages* (<http://pdf-images.sourceforge.net/>) that is available for the Linux, Mac OS X and Windows operating system platforms. On Linux, the tool is available as part of the *poppler* library which originated from the *xpdf*, an open-source viewer for X Windows graphical user interface system. The tool interprets the input PDF file and exports embedded images in commonly known PBM, PPM, or JPEG file formats. The images extracted in other formats can be converted to the JPEG format using ImageMagick® (<http://www.imagemagick.org/script/index.php>), which is an open-source command-line utility for Linux. The extracted high-resolution images converted to JPEG format are ready for identification and labeling by comparing them with the low-resolution labeled figure images available in the PMC bundle.

### 3.3 Identifying and Labeling the Extracted Images

As mentioned before, we have used two commonly known methods for image matching with some heuristic criteria for improving the matching. One method is based on image intensity projections on the coordinate axis, and the other is based on normalized cross-correlation.

#### 3.3.1 Image Identification based on Intensity Projection

We compared each extracted image with every image in the set of images supplied with the article bundle. The best matching image from the supplied bundle was selected for labeling when all of the following conditions were met:

- **Positive Gain Criterion**: This step is designed to ensure that using the extracted image would add value to the overall image indexing process. The extracted image must have a higher resolution than the image supplied with the article bundle and the ratios of image widths and heights must exceed a predefined threshold,  $TH_0$ .
- **Color Criterion**: Both, the extracted image and the image supplied with the article bundle must have the same color or intensity distribution.

- **Aspect Ratio Criterion:** The compared images must have equivalent aspect ratios. The difference in aspect ratios of the compared images is required to be less than a predefined threshold  $TH_1$ , for an image to qualify as a candidate for similarity matching.

**Similarity Measure Using Image Intensity Projection:** If all of the above conditions were satisfied, the extracted high-resolution image was downscaled to match the resolution of the supplied image under consideration and the image similarity between them was computed. In image comparisons based on Image Intensity Projections, the intensity projections of the images on the vertical and horizontal axes were compared by measuring the amount of overlap between the corresponding projections of the two images<sup>8</sup>. A match was registered if the amount of overlap exceeded a predefined threshold,  $TH_2$ .

For two grayscale images  $I_1$  and  $I_2$  that have the same dimensions, first, the intensity projections of the images in the vertical and horizontal directions were calculated. Then, the similarity of both projection distributions was measured using the average of the Bhattacharyya coefficient,

$$BC(I_1, I_2) = \frac{1}{2} \sum (\sqrt{p_1(x)p_2(x)} + \sqrt{q_1(y)q_2(y)})$$

where  $p_1(x)$  and  $p_2(x)$  are the horizontal projections,  $q_1(y)$  and  $q_2(y)$  are the vertical projections of image  $I_1$  and  $I_2$ , respectively, and  $x$  and  $y$  are the histogram bins of the projection vectors. Figure 3 shows example images (i, ii, and iii) and their vertical (iv), and horizontal (v) projection profiles.

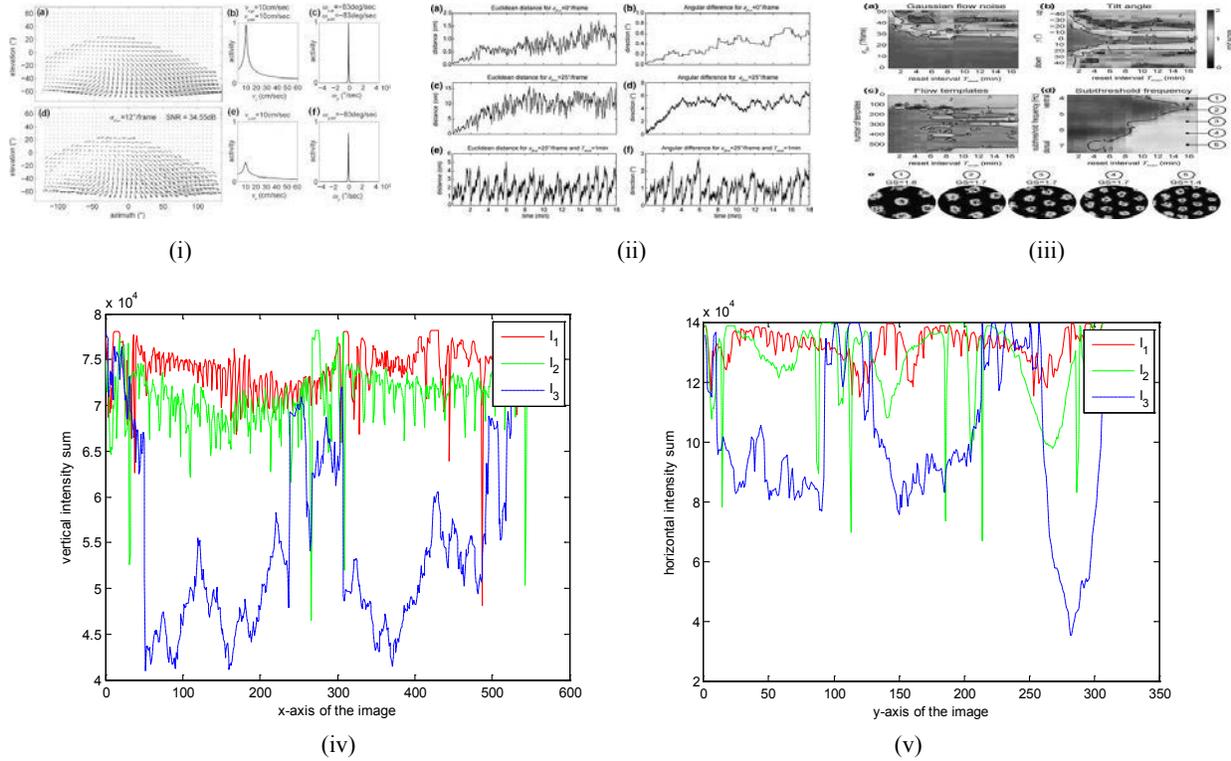


Figure 3. (i), (ii), and (iii) are images to be compared using the Image Intensity Projection method; and intensity profiles of three images on the x-axis (iv) and y-axis (v).

### 3.3.2 Image Identification based on Normalized Cross-Correlation

As in the first method, we compared each extracted image with every image in the set of images supplied with the article bundle. The best matching image from the supplied bundle was selected for labeling when the following conditions were met:

- **Aspect Ratio Criterion:** The compared images must have equivalent aspect ratios. The difference in aspect ratios of the compared images is required to be within a predefined threshold,  $TH_{AR}$ , before additional comparison tests can be performed.
- **Positive Gain Criterion:** This step is designed to ensure that using the extracted image would add value to the overall image indexing process. The extracted image must have a higher resolution than the image supplied with the article bundle. Since the aspect ratios are equivalent, greater image-width implies higher resolution. Unlike for the image intensity projection method, no minimum threshold or magnification ratio is required.

**Similarity Measure Using Normalized Cross-Correlation:** If the above conditions were satisfied, the two images were compared for similarity using *ImageMagick's* implementation of the Normalized Cross-Correlation (NCC) algorithm. The extracted image was downsampled to match the resolution of the bundled image. A match was registered if the value of the similarity measure exceeded a predefined threshold,  $TH_{NCC}$ .

Normalized Cross-Correlation is a technique commonly used for template matching and provides the correlation, or a measure of similarity between the compared images<sup>9</sup>. When using this technique, the image pixel intensities of the images are normalized prior to performing the actual image comparison. This step makes the image comparison invariant to linear brightness and contrast variations. The similarity score,  $S$ , is computed as:

$$S_{(I,R)} = \frac{1}{n} \sum_{(x,y)} \frac{(I_{(x,y)} - \mu_I)(R_{(x,y)} - \mu_R)}{\sigma_I \sigma_R}$$

where images  $I$ , and,  $R$  are being compared. Each image has  $n$  pixels distributed over  $x$ , and  $y$  dimensions. The mean and standard deviation of the image are expressed with  $\mu$  and  $\sigma$ , respectively. The resulting correlation is confined in the range [-1 to 1]. Figure 4 shows an example comparison between images using the NCC method.

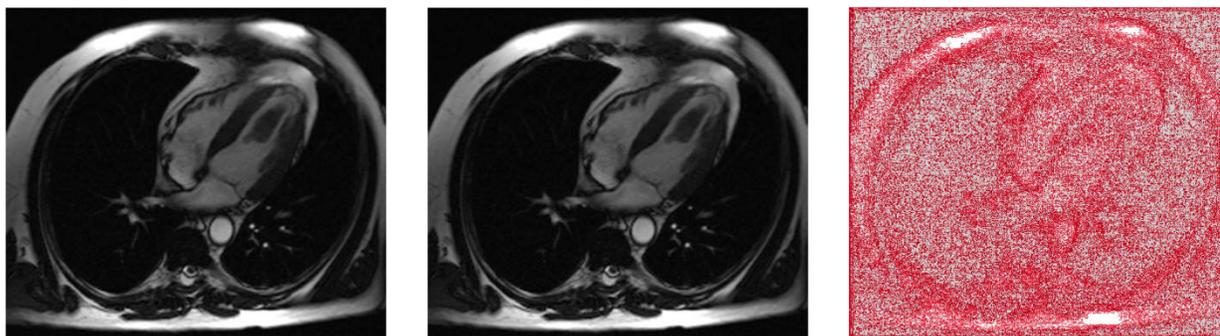


Figure 4. Example image matching using Normalized Cross-Correlation: (left) image supplied with the article bundle, (middle) extracted image downsampled to match the resolution of the supplied image, (right) figure showing the pixel positions where the intensities of the images differ. Although the absolute value of the pixel intensities for the above images differ for a large number of pixels, the overall intensity variation for the two images is very low, hence the similarity value of 99.977% using ImageMagick's compare function employing the NCC algorithm.

## 4. RESULTS

Our experiments were conducted on a set of 500 biomedical articles downloaded as compressed bundles from PubMed Central's ftp data feed (<ftp://ftp.ncbi.nlm.nih.gov/pub/pmc>). Each bundle contains a folder for the associated article. The folders contain the XML representation of the article; the standard (Web) resolution images; a PDF of the article (in most cases), and any ancillary or supporting documents. For this study, the PDF documents were used as source documents to extract the high-resolution versions of biomedical images. The standard resolution images were used for labeling the PDF-extracted images. In contrast to 1,348 standard resolution images, the image extraction utility extracted 1,590 high-resolution images from 351 PDF documents included in the bundles. In addition to the desired biomedical figures, the extracted set included "noise" images, such as publisher banners, borders and bullets, and spacers. For purposes of evaluation, the ground truth for the extracted images was generated by comparing them with those supplied

with the bundles and then assigning appropriate labels manually. Since some of the extracted PDF images did not contain all the elements of the corresponding images from the article bundle, we carefully examined each extracted image and a match was considered only when the extracted image contained all the elements of the corresponding image, and had a higher resolution. In total, we found 505 such extracted images. Our approach was evaluated on this ground truth set of manually labeled extracted images from the above set of 351 PDF documents. Tests were performed by setting the threshold values shown in Table 1.

Table 1. Threshold values used for image labeling using the methods described above.

<b>THRESHOLD VALUES FOR LABELING IMAGES</b>			
<b><u>Image Intensity Projection</u></b>		<b><u>Normalized Cross-Correlation</u></b>	
$TH_0$	1.25	$TH_{AR}$	0.05
$TH_1$	5	$TH_{NCC}$	0.70
$TH_2$	0.8		

Table 2 shows the results obtained by using the two labeling techniques described in this paper. Using Image Intensity Projection technique, we were able to correctly label 415 high-resolution images extracted from the PDF documents. Also, the overall precision and recall for this technique were 92.84% and 82.18%, respectively. Similarly, using the Normalized Cross-Correlation technique we were able to correctly label 408 high-resolution images extracted from PDF documents supplied with the biomedical articles. The precision and recall values achieved using this method were 84.30% and 80.79%, respectively.

Table 2. Results obtained using labeling techniques based on Image Intensity Projection and Normalized Cross-Correlation.

<b>RESULTS</b>			
<b><u>Image Intensity Projection</u></b>		<b><u>Normalized Cross-Correlation</u></b>	
Ground Truth Labeled Images	505	Ground Truth Labeled Images	505
Number of Images Labeled	447	Number of Images Labeled	484
Correctly Labeled Images	415	Correctly Labeled Images	408
Precision	92.84%	Precision	84.30%
Recall	82.18%	Recall	80.79%

We also performed Wilcoxon signed-rank test<sup>10</sup> on the results we achieved using the two described methods and found that either technique used for identification and labeling of images yield promising results and that the difference in the outcomes of the two techniques is statistically insignificant. Table 3 summarizes the results of Wilcoxon signed-rank test on the outcomes.

Table 3. Findings of the Wilcoxon signed-rank test on the results.

<b>WILCOXON SIGNED-RANK TEST</b>	
<b><u>Outcome</u></b>	
W+	1804
W-	1517
N	81
p	<= 0.5008

## 5. DISCUSSION

Our results show that both the techniques used for identifying and labeling the extracted high-resolution images rendered promising results. However, there were several challenges. When validating the images extracted from the source PDF documents we observed that PDF documents containing images with multiple layers do not yield images containing all the elements of the embedded images upon extraction using the tools used by us. Although some images extracted from

the PDF documents were higher resolution versions of the figures, they were missing several elements of the images that appeared in print and on the standard resolution versions. This is a common problem in extraction of these images and details of a published method<sup>5</sup> that addresses the problem are lacking. Analysis of the outcome of the techniques described above showed that a few of these images were also reported as correct matches although most of them are correctly filtered out. Figure 5 shows an example pair of matched images where the standard resolution image contains text elements missing in the corresponding high-resolution version extracted from the supplied PDF document. Images missing a few elements were also found when the experiments were run using stricter values for the thresholds. Upon checking the source PDF documents, we found that the images contained all elements when viewed in a PDF viewer. The majority of false positive images contained some text along with the underlying graphic content. We also observed that the text contained within such images was recognized by the PDF readers and was part of the searchable content. Thus, we concluded that either the PDF was subjected to processing or optimizations like Optical Character Recognition (OCR), or the images were embedded as layers of text super-imposed on, but not fused with, the image pixel plane.

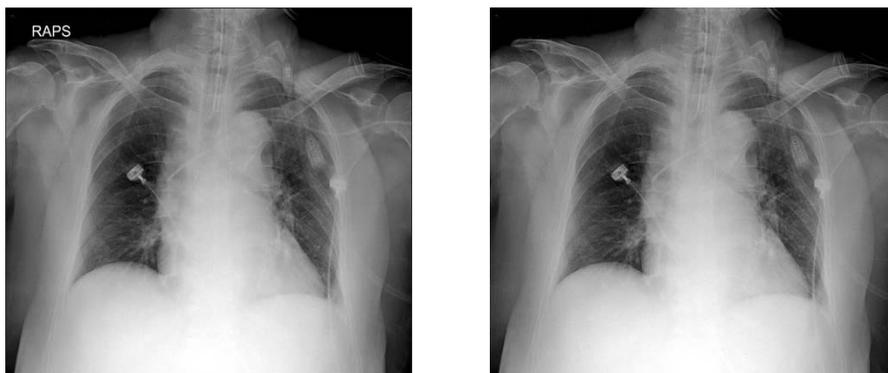


Figure 5. Missing elements in the images extracted from PDF documents. Text ‘RAPS’ seen in the supplied image (left) is missing in the high-resolution version (right) of the image extracted from the bundled PDF document. Images rescaled to have similar size.

## 6. CONCLUSION AND FUTURE WORK

In this paper, we propose a method to utilize the PDF documents as a source of high-resolution version of the figures downloaded from PubMed Central. The process was developed to serve the needs of our Open-I medical article retrieval system. The major challenge of this task is recognizing that an image extracted from a PDF document is not a whole figure. For example, textual labels may be missing. Therefore, the approach needs to capture such important content difference while at the same time allow certain differences that are acceptable, such as slight contrast differences and slight translation. We employed two methods aiming at this goal. We tested our approach on a ground truth dataset of 500 articles which contains 1348 figures and 1590 PDF extracted images. We achieved a precision of 92.84% with a recall of 82.02%. Both methods yielded promising results and the differences between the methods were not statistically significant.

A possible extension to the work described in this paper could be to merge the identification and labeling techniques described in this paper to improve the accuracy and further reduce the number of false matches. Optical character recognition techniques could also be utilized to analyze the extracted images after they have been labeled to detect missing elements and to flag the image suitably. Another desirable extension would be developing a technique for extraction and reconstruction of complex, multi-layered graphics and figures from PDF documents.

## ACKNOWLEDGMENTS

This work was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and the Lister Hill National Center for Biomedical Communications (LHNCBC).

## REFERENCES

- [1] Demner-Fushman, D., Antani, S. K., Simpson, M., Thoma, G. R., "Design and development of a multimodal biomedical information retrieval system," *Journal of Computing Science and Engineering*, 6(2), 168-177 (2009).
- [2] You, D., Rahman, M. M., Antani, S. K., Demner-Fushman, D., Thoma, G. R., "Text- and content-based biomedical image modality classification, " *Proc. SPIE Medical Imaging*. Orlando, FL., 8674-21 (2013)
- [3] "ISO 32000-1:2008 - Document management -- Portable document format -- Part 1: PDF 1.7" Iso.org, 2008-07-01.
- [4] Rahman, M. M., Antani, S. K., Thoma, G. R., "A Query Expansion Framework In Image Retrieval Domain Based On Local and Global Analysis," *Information Processing and Management*, 47(5), 676-691 (2011).
- [5] Xu, C., Tang, Z., Tao, X., Shi, C., "Graphic composite segmentation for PDF documents with complex layouts," *Document Recognition and Retrieval XX, Proceedings of SPIE 8658*, 86580E1-86580E10 (2013)
- [6] Shao, M., Futrelle, R., P., "Recognition and classification of figures in PDF documents," in *Graphics Recognition. Ten Years Review and Future Perspectives*. LNCS, 239-251 (2006).
- [7] Lin, X., Gao, L., Tang, Z., Lin, X., Hu, X., "Mathematical formula identification in PDF documents," 2011 *International Conference on Document Analysis and Recognition*, 1419-1423 (2011)
- [8] Herman, G. T., "Fundamentals of computerized tomography: Image reconstruction from projections," 2<sup>nd</sup> ed., Springer, ISBN 978-1-85233-617-2 (2009)
- [9] Lewis, J. P., "Fast normalized cross-correlation," *Industrial Light & Magic* (1995)
- [10] Wilcoxon, F., "Individual comparisons by ranking methods," *Biometrics Bulletin*, 1(6), 80-83 (1945)