# An Automatic System for Extracting Figures and Captions in Biomedical PDF Documents

Luis D. Lopez, Jingyi Yu, Cecilia N. Arighi, Hongzhan Huang, Hagit Shatkay, Cathy Wu
*Department of Computer and Information Sciences*
*University of Delaware. Newark, USA*
*ldlopez@udel.edu, yu@eecis.udel.edu, arighi@dbi.udel.edu, huang@dbi.udel.edu, shatkay@cis.udel.edu, wuc@dbi.udel.edu*

***Abstract*—Figures in biomedical articles often constitute direct evidence of experimental results. Image analysis methods can be coupled with text-based methods to improve knowledge discovery. However, automatically harvesting figures along with their associated captions from full-text articles remains challenging. In this paper, we present an automatic system for robustly harvesting figures from biomedical literature. Our approach relies on the idea that the PDF specification of the document layout can be used to identify encoded figures and figure boundaries within the PDF and enforce constraints among figure-regions. This allows us to harvest fragments of figures (subfigures), from the PDF, correctly identify subfigures that belong to the same figure, and identify the captions associated with each figure. Our method simultaneously recovers figures and captions and applies additional filtering process to remove irrelevant figures such as logos, to eliminate text passages that were incorrectly identified as captions, and to re-group subfigures to generate a putative figure. Finally, we associate figures with captions. Our preliminary experiments suggest that our method achieves an accuracy of $95\%$ in harvesting figures-caption pairs from a set of $2,035$ full-text biomedical documents from BioCreative III, containing $12,574$ figures.**

***Keywords*-images; figures and captions; information retrieval; biomedical documents; biomedical images;**

## I. Introduction

Figures in the biomedical literature provide a unique source of information. They may contain useful details about experimental settings, methodology, and procedures to help readers better comprehend the contents and interpret the results. Figures, along with their associated captions, can effectively illustrate hypotheses and highlight the contributions stated in scientific publications.

Despite the usefulness of figures, they are not readily accessible in public databases. The vast size of the biomedical literature makes essential the use of automated systems to robustly harvest figures from biomedical publications.

Although most of current information retrieval methods within the biomedical domain utilize only the text when searching for relevant articles, recent work has indeed started utilizing information from figures and captions in the context of biomedical information retrieval [1], [2]. There is also an emerging trend of coupling figures and text (especially figure captions) for biomedical literature mining. Examples include work by Murphy *et al* [3] on identifying documents that contain information and images relevant to protein subcellular localization, by Shatkay *et al* on the integration of text and images for biomedical document categorization [4], and recent work by Demner-Fushman *et al* [5] that uses figure captions to help classify, archive and retrieve various types of CT/MRI images. While their system automatically extracts figures from documents, it requires some manual intervention to extract figure captions and associate them with the respective figures. Further manual intervention is needed for correcting some of the errors, such as merging image-panels into complete figures, and removing images that are not content-bearing.

In this paper, we present an automatic system for robustly harvesting figures from the biomedical literature. A unique feature of our system is its capability to simultaneously extract figures and locate their associated captions. In particular, we intended to address two fundamental challenges. First, to remove figures that are not part of the publication (e.g. logos). Second, to identify and fix fragmented figures that are stored separately in PDF documents.

We demonstrate the utility of our method by applying it to a corpus obtained from the BioCreative III, Interaction Method task (IMT) [6]. This large, public dataset consists of $2,034$ full-length biomedical articles describing experimental techniques for studying PPI, thus containing a large variety of biomedical images. Furthermore, the IMT challenge itself represents an important biomedical application for image mining. Our results show that we correctly extracted $95\%$ figure-caption pairs from the dataset, which consists of over $12,500$ figures.

## II. Our Approach

A high-level view of our pipeline for automatically extracting figures along with their associated captions from biomedical documents is shown in Figure 1. Our system consists of four key components: (A) a **PDF Operator Parser** that simultaneously recovers figures and captions from each PDF document, (B) a **Figure Filter** that identifies and removes “noise figures”, such as journal logos and fixes fragmented figures (e.g. merging subfigures), (C) a **Caption**

978-0-7695-4574-5/11 $26.00 © 2011 IEEE
DOI 10.1109/BIBM.2011.26

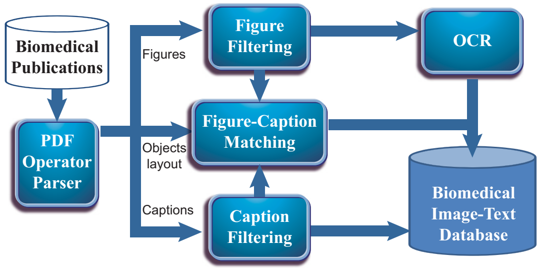

Figure 1. Overview of our system to harvest figures and captions from biomedical publications

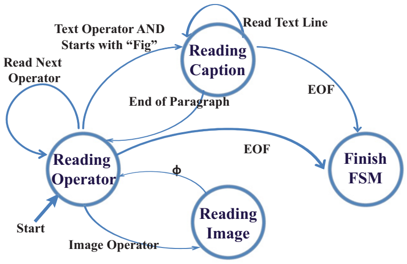

Figure 2. State transition diagram of the Finite State Machine (FSM) implemented to recover information from PDF files

**Filter** that evaluates and removes the captions incorrectly harvested in (A), and (D) a **Figure-Caption Matcher** that associates the recovered figures and captions. In addition, we store the resulting figures and their associated captions in a local database and link them with other important information (e.g PMID, XML files, and PDF files) for potential future use.

### A. *PDF operator parser*

A PDF document consists of a set of operators describing the text and graphic objects to be displayed. Each operator has specific parameters to define layout and formatting options [7]. To open the PDF files and extract these operators, we use a modified version of Xpdf (http://foolabs.com/xpdf/), a public domain tool. As shown in figure 2, to simultaneously recover captions and figures from the set of PDF operators, our solution uses an event-driven Finite State Machine (FSM) model with four states.

The initial state is the "**Reading Operator**" that simply reads the set of operators generated by the Xpdf tool. We consider all the paragraphs starting with "Fig" as potential captions, so when this state finds a text operator with its string starting with "Fig" or any variant (e.g. "FIG") we create a transition to the "Reading Caption" state. Similarly when this state finds a graphics operator defining a figure, it transits to the "Reading Image" state.

The "**Reading Caption**" state creates a new string of text that contains a potential caption. In PDF files, it is common that a single paragraph can be split across multiple operators. Therefore, we consider the subsequent lines of text as part of the current caption until we identify the end of the paragraph. A new paragraph or the end of the file indicates a transition out of the "**Reading Caption**" and induces a transitions to either the "**Reading Operator**" or "**Finish FSM**" states, respectively.

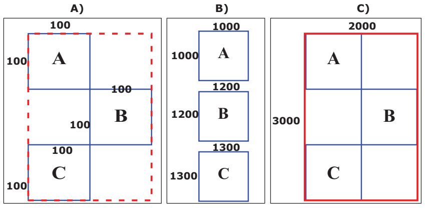

Figure 3. Rendering a high resolution figure from a set of subfigures. (A) Whole figure layout (consiting of three subfigures, denoted A,B and C, in the low resolution page space. (B) Actual subfigure dimensions. (C) Reconstructed high resolution figure.

The "**Reading Images**" state retrieves the image information and its pixel values, stores the image in an internal data structure for further processing and returns the control to the "**Reading Operator**" state. Finally, the "**Finish FSM**" state simply ends the recovering process.

### B. *Figure Filter*

Once our FSM harvests all the figures, the latter are filtered to remove non-informative figures such as logos attached by publisher. To do so, we first analyze the document layout information stored in the PDF file, such as figure and text margins, column widths, and line spaces. We found that logos are found outside the text margins, then, for each extracted figure we evaluate its position with respect to the rectangle defined by the exterior text margins and exclude figures that lie outside this area.

Another challenge is the lack of uniform standards for embedding figures in biomedical documents. Ideally, each figure in a publication should correspond to a single figure in the PDF file. However, in practice a figure is often fragmented and stored as a set of subfigures. To resolve this problem, we group all the consecutive operators defining figures preceding each caption into one figure, and render its corresponding full-resolution image. We use the document layout to compute subfigures size in the low resolution page dimensions. Next, for each subfigure, we calculate its scaling. Finally, to avoid undersampling we generate the final figure using the smallest scaling factor obtained over for all the constituents subfigures, and map the pixel values of each input image to the final image. Figure 3 shows an example of this process.

### C. *Captions Filter*

The FSM described in Section II-A correctly identified all captions. However, it may also mistakenly identify references to figures occurring within the text as captions. Therefore, we evaluate the recovered information to first identify and then remove references to figures. For every string of text in the potential caption set, we create a descriptor using only the first and second words in the string. The descriptor preserves the alphabetical characters, special characters, and punctuations from both words. In the

Table I
EXAMPLE OF THE CONSTRUCTION META DESCRIPTORS

| Caption | Meta descriptor | Group |
|---|---|---|
| Figure 1. ... | Figure#. | 1 |
| Fig. 1 ... | Fig.# | 2 |
| Fig. 2 ... | Fig.# | 2 |
| Figure 10. ... | Figure#. | 1 |

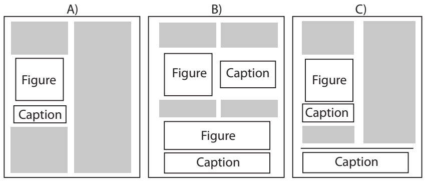

Figure 4. Three categories of the figure-caption matching problem. (A) 1-to-1 matching. (B) N-to-N matching. (C) N-to-M matching.

descriptor we substitute the numbers by the special symbol $'\#'$. We then cluster strings sharing identical descriptors, and use the integer value in the second word to calculate the number of unique subfigures linked in each group. Table I shows an example construction of descriptors from a set of string of captions. Finally we select the group with a maximum number of unique links to subfigures and discard the rest.

### D. Figure-Caption Matcher

Once we extract the figures and the captions, we associate each figure $f_i$ with a caption $c_j$, our goal is to use geometric and structural cues to compute the optimum match between the corresponding objects. Our approach separately handles, 1-to-1, N-to-N, and N-to-M matching problems, which correspond to associate 1 figure to 1 caption, $N$ figures to $N$ captions and $N$ figures to $M$ caption, respectively, as shown in Figure 4. Given the set of figures and captions we classify them according to their structural information into left column, right column and double column object. For each non-empty page we first compute the association cost between all figure-caption pairs across the page.

For a given pair of objects $f_i$ and $c_j$ we compute their matching cost $C(f_i, c_j)$ as the minimum distance $d_{f_i,c_j}$ between their corresponding boundaries multiplied by a penalty cost $p_{f_i,c_j}$ that prioritizes the matches between figures and captions with similar structural information $C(f_i, c_j) = d_{f_i,c_j} * p_{f_i,c_j}$. If $f_i$ and $c_j$ are objects in the same column, the penalty cost is set to be 1; otherwise it is set to be 10.

Once we finish building the cost matrix we start matching the objects. In the base case **1-to-1 matching** we simply associate the figure to the caption in the page.

The **N-to-M matching** and **N-to-N matching** problems are solved by applying a greedy algorithm to find the optimal global association. We start by finding the figure-caption pair with the minimum value in the matching cost table. Next, we associate them and recalculate the cost matrix excluding the previously associated objects. We repeat this process until we exhaust the figures or caption sets from the current page. Finally we associate the unmatched objects by repeating the previous steps including objects from different pages, and recomputing the matching cost by including the disparity between their pages in the original cost equation $C(f_i, c_j) = d_{f_i,c_j} * p_{f_i,c_j} * disparity_{f_i,c_j}$.

Similar to previous approaches [1], [8], we extracted text embedded inside figures to enable indexing figures based on this text. We use the commercial ABBYY OCR software for recognizing characters in high resolution figures recovered in the previous steps. We recover the text within the subfigures directly from the PDF file. The OCR software correctly recognize English characters, but its performance is lower when extracting Greek and other special symbols. Finally, we use a standard MySQL database to store each figure, and corresponding subfigure captions, and embedded text. We then associate the figures with other publicly available information (PDF files, XML files, PMID). We also use both the text embedded in the figures and the words from the captions as keywords for indexing and retrieval.

## III. RESULTS AND DISCUSSION

To evaluate the system performance we have collected $2,035$ full-text biomedical documents from the corpus provided by BioCreative III for the Protein-Protein Interaction (PPI) Interaction Method Task (IMT) [6]. It is important to note that IMT provides a variety of document formats (XML, PDF, TXT, and HTML), and we use the PDF format as this is the only set to contain all the figures. In our experiment, we were able to generate a dataset consisting of $12,574$ figures with associated captions.

As mentioned in Section II-B, an important feature of our system is that it can automatically remove journal logos. In the Biocreative III corpus: about $30.4\%$ of figures are logos. In traditional figure extraction solutions, these irrelevant figures not only incur computational processing overhead but also make it difficult to reliably associate figures with captions. In contrast, our solution detects and removes the logo figures before associating figures with captions. In our experiment $5,832$ figures were identified as logos using our algorithm. Of these, $5,830$ were true logos. Notably, no logos were missed by our system, only two incorrect labeling of non-logo figures (false-positives) occurred; these non-logo figures were initially defined outside the text margin and were then rotated and translated back inside the margin, and were therefore captured by our system.

For the task of merging subfigures into figures, our system identified $40,604$ subfigures, and merged them into $2,158$ figures. Of these, $2,149$ were correctly merged using our algorithm. Figure 5 shows two examples of the merged subfigures. The first row shows our result on a figure obtained from article PMID:16096643 [9] that is composed of multiple subfigures. Column A shows a figure generated using our method as explained in Section III. Column B shows the manually extracted figure containing the corre-

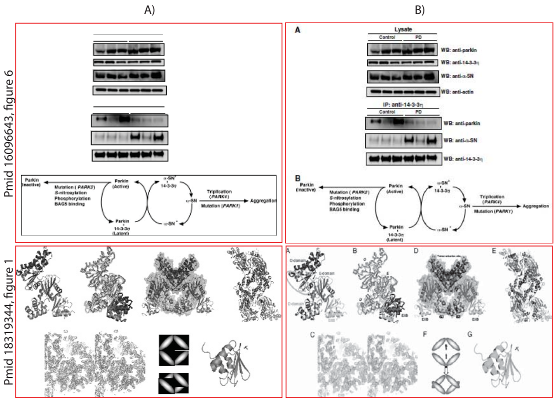

Figure 5. Figures generated by our system through merging subfigures. (Figures generated by our system (A) vs. original figures (B).

sponding subfigures. It is important to note that the quality of the figure produced by our system is typically higher than that of the one manually extracted. This is because when a figure is extracted using tools, such as screen capture, it is often stored at a lower resolution [1]. Our method keeps the original resolution and then merges the subfigures at a user-defined resolution that can be dynamically adjusted.

The second row of Figure 5 shows a more challenging example of a figure composed by thousands of heterogeneous subfigures from PMID:18319344 [10]. each with a different opacity. These subfigures provide no biological meaning unless they merged together. Since our approach does not attempt to recover the opacity value for each subfigure, our merged result (Column A) slightly differs from the ground truth (Column B).

Finally, to evaluate the performance of our figure-caption matcher we used our system to obtain caption-figure pairs and verified the results manually. Out of $12,574$ pairs, our system correctly matched $12,069$, which is $95.98\%$. However, our algorithm cannot currently handle well complex structures combined with inconsistent layouts. For example, a figure and its caption may be separated between different pages (e.g., Figure 2 in PMID:16362034 [11]), a single caption can be associated with multiple figures (e.g, Figure 1 in PMID:19303849 [12]), figures and their captions can be ordered inconsistently (e.g., Figures 1,2, and 3 in PMID:16239925 [13]). While our system cannot handle these cases, it can recognize them, and notify the user to manually handle the case.

**Concluding Remarks.** We have presented a new automatic system for harvesting and associating figures and captions from biomedical publications in PDF format. The approach utilized constraints derived from the document layout to guide the correct combination of subfigures into figures and the figure-caption matching process. Specifically, we have developed filters that identify and remove irrelevant figures, such as logos and equations, harvests putative subfigures and captions, groups the subfigures into their original figures and associates them with their matching captions using geometric and structural cues from the document layout. We have applied our system to a large, publicly available biomedical corpus, demonstrating that our system automatically extracts figures from the PDF files and associates them with the respective captions, both with a very high level of performance.

## REFERENCES

[1] D. Kim and H. Yu, “Figure Text Extraction in Biomedical Literature,” *PLoS ONE*, vol. 6, no. 1, pp. e15 338+, Jan. 2011.

[2] N. Chen, H. Shatkay, and D. Blostein, “Use of figures in literature mining for biomedical digital libraries,” in *Proc. of the 2nd IEEE Int. Conf. DIAL*, 2006, pp. 180–197.

[3] R. F. Murphy, M. Velliste, and G. Porreca, “Robust classification of subcellular location patterns in fluorescence microscope images,” in *Procs. of the 12th IEEE Workshop on Neural Networks for Signal Processing*, 2002, pp. 67–76.

[4] H. Shatkay, N. Chen, and D. Blostein, “Integrating image data into biomedical text categorization,” in *Intelligent Systems in Molecular Biology*, vol. 22, 2006, pp. 446–453.

[5] D. Demner-Fushman, S. Antani, M. Simpson, and G. R. Thoma, “Annotation and retrieval of clinically relevant images,” *Int. J. of Med. Inf.*, vol. 78, pp. 59–67, 2009.

[6] M. Krallinger *et al*, “The protein-protein interaction tasks of biocreative iii: classification/ranking of articles and linking bio-ontology concepts to full text,” *BMC Bioinformatics*, 2011.

[7] C. Adobe Systems Inc, *PDF Reference with Cdrom*, 2nd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2000.

[8] S. Xu, J. Mccusker, and M. Krauthammer, “Yale image finder (yif): a new search engine for retrieving biomedical images,” *Bioinformatics/computer Applications in The Biosciences*, vol. 24, pp. 1968–1970, 2008.

[9] S. Sato, T. Chiba, E. Sakata, K. Kato, Y. Mizuno, N. Hattori, and K. Tanaka, “14-3-3eta is a novel regulator of parkin ubiquitin ligase.” *EMBO J*, vol. 25, 2006.

[10] T.-W. Nam, H. I. Jung, Y. J. An, Y.-H. Park, S. H. Lee, Y.-J. Seok, and S.-S. Cha, “Analyses of mlc-iibglc interaction and a plausible molecular mechanism of mlc inactivation by membrane sequestration.” *Proc Natl Acad Sci USA*, vol. 105, no. 10, pp. 3751–6, 2008.

[11] F. Qiao, B. Harada, H. Song, J. Whitelegge, A. Courey, and J. Bowie, “Mae inhibits pointed-p2 transcriptional activity by blocking its mapk docking site.” *EMBO J*, vol. 25, no. 1, pp. 70–9, 2006.

[12] R. H. F. Wong, I. Chang, C. S. S. Hudak, S. Hyun, H.-Y. Kwan, and H. S. Sul, “A role of dna-pk for the metabolic gene regulation in response to insulin.” *Cell*, vol. 136, no. 6, pp. 1056–72, 2009.

[13] S. C. Sampath, R. Ohi, O. Leismann, A. Salic, A. Pozniakovski, and H. Funabiki, “The chromosomal passenger complex is required for chromatin-induced microtubule stabilization and spindle assembly,” *Cell*, vol. 118, pp. 187–202, 2004.