A novel figure panel classification and extraction method for document image understanding

Xiaohui Yuan* and Dongyu Ang

University of North Texas,
Denton TX 76201, USA
E-mail: xiaohui.yuan@unt.edu
E-mail: dongyu.ang@unt.edu
*Corresponding author

Abstract: With the availability of full-text documents in many online databases, the paradigm of biomedical literature mining and document understanding has shifted to analysis of both text and figures to derive implicit messages that are unforeseen with text mining only. To enable automatic, massive processing, a key step is to extract and parse figures embedded in papers. In this paper, we present a novel model-driven, hierarchical method to classify and extract panels from figures in scientific papers. Our method consists of two integrated components: figure (or panel) classification and panel segmentation. Figure classification evaluates each panel and decides the existence of photographs and drawings. Mixtures of photographs and non-photographs are divided into subfigures. The splitting process repeats until no further panel collage can be identified. Detection of highlighted views is addressed with Hough space analysis. Using reconstruction from Hough peaks, enclosed panels are retrieved and saved into separate files. Experiments were conducted with a total of 360 figures extracted from two sets of papers that are retrieved with difference sets of keywords. Experimental results demonstrated that our method successfully segmented figures and extracted photographs and non-photographs with high accuracy and robustness. In addition, our method was able to identify zoom-in views that are superimposed on the original photographs. The efficiency of our method allows online implementation.

Keywords: classification; image segmentation; document analysis; literature mining.

Reference to this paper should be made as follows: Yuan, X. and Ang, D. (2014) 'A novel figure panel classification and extraction method for document image understanding', *Int. J. Data Mining and Bioinformatics*, Vol. 9, No. 1, pp.22–36.

Biographical notes: Xiaohui Yuan received a BS in Electrical Engineering from Hefei University of Technology, China in 1996 and a PhD in Computer Science from Tulane University, USA in 2004. After graduation, he worked at the National Institutes of Health for two years on medical image analysis. He is an Assistant Professor in the Computer Science and Engineering Department at the University of North Texas. His research interests include data mining, machine learning, image processing and pattern recognition and he has published more than 50 technical papers. He is a member of IEEE and SPIE.

Dongyu Ang received a BS in Computer Science from the Hefei University of Technology, China in 2007. He is currently pursuing his PhD in the Department of

Computer Science and Engineering at the University of North Texas. His research interests include data mining and image processing.

1 Introduction

It is well-recognised that "a picture is worth a thousand words". The critical role of figures in understanding the contents of scientific documents and the limited number of studies necessitate the development of automatic panel extraction and classification methods. Much current work in biomedical literature mining aims at extracting information and discovering knowledge for populating biomedical digital libraries. Most existing methods focus on the analysis of text, typically from abstracts, to perform tasks, such as document classification, named entity tagging and information extraction (e.g., extracting interactions between proteins). With abundant full-text documents made available in online databases, the paradigm of biomedical literature mining has shifted to combine the analysis of both text and embedded figures to derive implicit messages that are unforeseen through text mining or image mining only (Samuel et al., 2010).

Unlike the traditional image mining problem, understanding figures embedded in scientific literature faces unprecedented challenges. Figures in papers are usually collages of multiple panels. These panels could be of the same type or different types, e.g., fluorescence images, statistical plots, analysis procedure diagrams, etc. To analyse these figures and extract information, a key step is to split figures into panels such that each one contains only a single image. Hence, the challenges arise from mixed types of images and various panel layouts (Lu anb et al., 2009). In addition, it is often used to illustrate details with highlighted views (e.g., superimposed zoom-in views of a sub-region) over the original image. These highlighted views are especially important and contain strong evidence discussed in the paper. However, correctly identifying and segmenting these panels are non-trivial and, yet, critical in the shifted paradigm of text and image-based literature mining.

Despite the great needs in automatic figure analysis, limited studies have been conducted to extract and analyse panels from papers. The existing methods rely heavily on heuristics. Murphy et al. (2001) and Qian and Murphy (2008) extracted figures from papers and segmented fluorescence microscopy images using projected image intensity histograms. Lu anb et al. (2009) presented an automatic figure categorisation method in the context of the digital library, which treats the collage of panels as a unit. Assumptions are made that panels are available for use. In Doermann (1998), indexing and retrieval of document images are discussed, which require prior knowledge of figures in the papers.

In this paper, we present a novel method to classify and extract panels from figures in biomedical papers. Our overarching goal is to make figure analysis an automatic process to enable text and image-based literature mining. While our focus in this paper is to identify photographs for further analysis, artificially generated graphs, such as plots and diagrams, are also expected to be correctly segmented and saved in a tree data structure. Our method consists of figure (or panel) classification and segmentation. Figure classification determines the existence of photograph in a figure. Gaussian models are constructed for photographs and non-photographs. Figures and panels are evaluated based on the models to determine

their categories. An iterative panel-splitting process follows the classification and continues until no further separation margins are identified in the subfigures.

Our contribution in this paper include a novel model-based method for panel classification that derives generalisable patterns from examples and a hierarchical panel extraction method that extract photographs, plots, diagrams and embedded highlighted views. Our method processes figures in an automatic fashion such that it can be integrated to online or off-line document understanding and literature mining.

The remainder of this paper is organised as follows. In Section 2, we review related works in the fields of literature mining and document image analysis. In Section 3, we present our method for automatic panel extraction. Our proposed new techniques for model-driven image classification, extracting panels and segmenting sub images, are discussed in detail. In Section 4, we present our experimental results and demonstrate the performance by comparing it to Murphy's method (Murphy et al., 2001). Correctness, robustness, efficiency are discussed with examples and statistical analysis. In Section 5, we conclude our paper and outline possible future extensions.

2 Related work

Fisher et al. (1990) proposed a rule-based system for segmenting a document image into text and non-text blocks. Rule-based systems used in the document image domain, however, have not fully exploited the depth and breadth of knowledge that is available about specific document domains. Murphy et al. (2001) and Qian and Murphy (2008) piloted a structured literature image finder system, within which they attempted to parse text and figures in biomedical literature. Splitting panels in figures is a key part of image analysis in their system, in which cumulative density along the x and y axes and the locations of density peaks were used to recursively split figures and subfigures into panels. Threshold selection and over segmentation are obstacles in the path of fully automating the process and achieving robustness. Lu anb et al. (2009) focus on automatically categorising figures in scientific journals. Their goal was to automate the efforts of extracting, categorising and indexing figures from scientific documents and to make use of information in figures to enhance the performance. An assumption was made in their studies that most of the panels in figures are extracted as standalone images. Shatkay et al. (2006) used figures to help classify biomedical literatures. They conducted a connected components analysis-based segmentation method. Such analysis is performed on thresholded black-and-white images, where connected components are regions of neighbouring foreground pixels. The connectedness is defined based on the eight neighbours of each pixel.

Document image understanding has also been studied in the context of digital libraries. It aims at analysis of image representation of a document, for instance, a digitised paper, using a scanner, into high-level semantic descriptions. Many works have been done in document image understanding, e.g., physical and logical structure analysis (Niyogi and Srihari, 1995; Mao and Rosenfeld, 2003) and indexing and retrieval of document images (Doermann, 1998). Niyogi and Srihari (1995) developed a computational model for document logical structure derivation. A rule-based control strategy was developed that utilises the data obtained from analysing digitised document images and makes inferences

using a multi-level knowledge base of document layout. Mao and Rosenfeld (2003) reviewed the previous methods on document physical layout representations, document logical structure representations and performance evaluation of document structure analysis algorithms. Document physical layout can be better represented with tree structures derived from a set of rules, as suggested in Yamashita et al. (1991), Tsujimoto and Asada (1990) and Fisher (1991). Such a tree structure gives a hierarchical view of the document contents and could assistant indexing and structural analysis. Recognising diagrams in documents has been discussed by Blostein et al. (2000) and several approaches were developed including blackboard systems, stochastic grammars, hidden Markov models and graph grammars. Diagram recognition techniques have been developed to address a great variety of scenarios in which diagram recognition is used. Most of these are research systems, or systems that are customised to address the needs of one particular client (Arias et al., 1993, 1998). Form processing was investigated by using hidden Markov models to detect parallel lines in a form (Zheng et al., 2005).

3 Panel segmentation and classification

Papers in biology and medicine use a combination of plots, diagrams and various photographs to illustrate key findings. Figure 1 shows two typical figures in published papers of our collection. It is very common that multiple panels (or subfigures) of different types are organised in layouts that are unstructured. Our goal in this paper is to separate the panels in the figure and extract the photographs and non-photographs for further analysis.

We categorise panels into photographs and non-photographs (i.e. plots, diagrams and other artificial drawings). In this paper, photographs are the illustrations generated from imaging devices, such as microscope images and fluorescence images, or some image processing techniques over the original images; whereas non-photographs are the artificial illustrations of analysis steps or outcomes. An example of non-photographs is shown in the right-top corner of Figure 1(a). In our previous studies on literature mining (Samuel et al., 2010), we notice that plots and diagrams are usually discussed rather thoroughly within the text. On the other hand, a few image properties are discussed that are aligned with the study objectives, while abundant image features are left undiscussed. Hence, analysing photographs and extracting complimentary information is greatly valuable in the path of achieving comprehensive literature mining.

Our method consists of two integrated components: figure (or panel) classification and panel segmentation. An overview diagram of our method is shown in Figure 2. The figure/panel classification decides if there exists a photograph in the figure or panel. The mixture of plots and photographs are divided into subfigures. The splitting process repeats until no further panel collage can be identified. In all photograph panels, detection of highlighted views (i.e. zoom-in views) is also done. Our method maps a photograph into the Hough space with a line template. Using reconstruction from the Hough peaks, the enclosed panels are retrieved and saved into separate files.

3.1 Figure/panel classification

To achieve figure/panel classification, our idea is to construct models of photographs and non-photographs from a set of training images and make prediction with these models.

Figure 1 Examples of typical figures in scientific papers and our expected results. (a) and (b) are figures in Drin et al. (2008) and Xie and Hua (2010), respectively. (c) and (d) show the expected panel segmentation results of examples in (a) and (b). Photographs are identified and saved into separate files (see online version for colours)

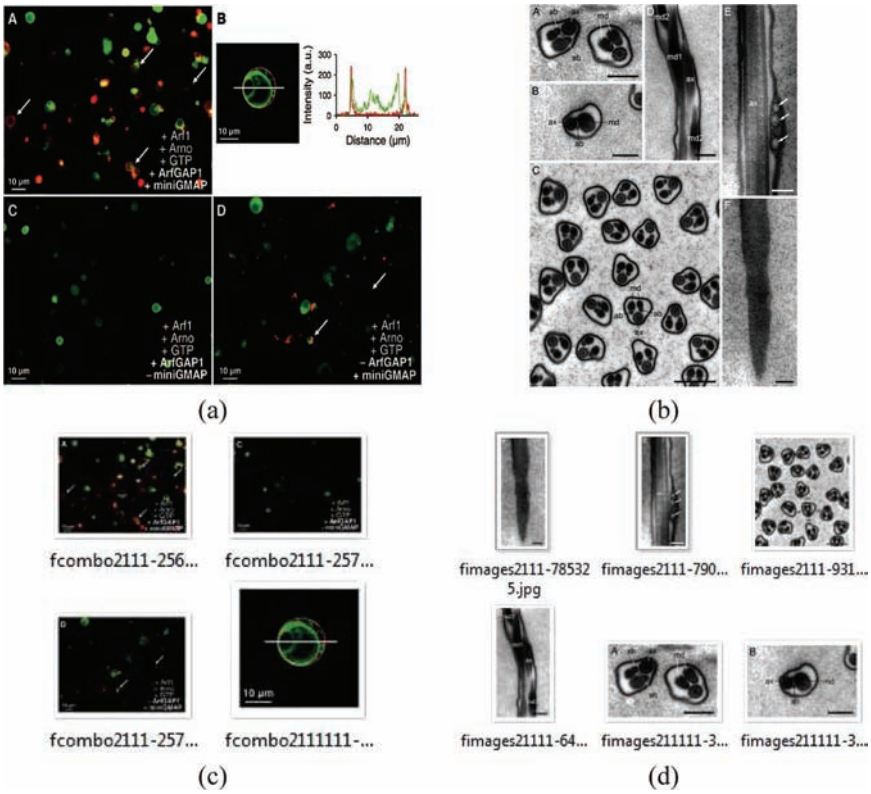
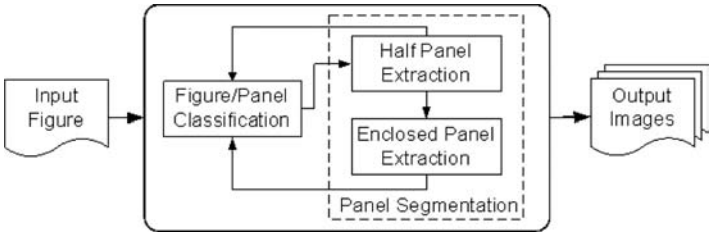


Figure 2 An overview of our method



Given examples of photographs and non-photographs, the normalised histograms of the images are modelled with multivariate Gaussian functions:

$$\begin{aligned} hI &= G(\bar{h}_I, \Lambda) \\ hP &= G(\bar{h}_P, \Lambda P) \end{aligned} \tag{1}$$

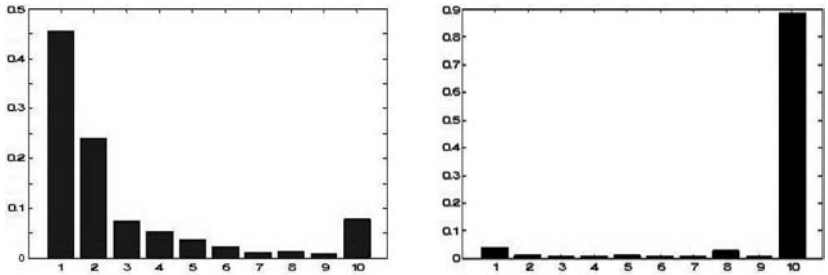
where h_I and h_P denote the histogram models of the photographs and plots, respectively. Function G denotes the Gaussian function with mean h and covariance A. Both h_I and h_P are normalised such that

$$\int h_I = 1 \quad \text{and} \quad \int h_P = 1.$$

The normalisation ensures that the histograms are invariant to figure size.

Figure 3 depicts aggregated normalised histograms of photographs and non-photographs. Ten quantisation bins were used in generating the histograms, which were derived from 54 and 36 examples for photographs and non-photographs, respectively. Figure 3(a) and (b) illustrate the average normalised histogram of photographs and non-photographs, respectively. It is clear that the average histograms of the two categories are significantly different.

Figure 3 Normalised histograms of photographs (a) and non-photographs (b)



(a) (b)

Note that our model is not profiling the histogram trend across the quantisation bins. It encodes the mean and variations of each bin. In this model, we assume independence between bins. Given 10 quantisation bins, as shown in Figure 3, the covariance matrix of each Gaussian model is a 10 by 10 matrix that contains only the diagonal values. In addition, it is necessary to have a model for each category of panels. Because it is common to have a collage of photographs and non-photographs in one panel, our classifier results in three decisions: photographs, non-photographs and combination.

To classify a figure (or a panel) x, the normalised histogram of x is computed, denoted with h_x and $\int h_x = 1$. Assume the histogram of x is quantised into M bins. The probability of x being a photograph can, hence, be calculated using two tailed z-test as follows:

$$p_I(x) = \frac{2}{M}\sum_{m=1}^{M} H\left(\frac{\bar{h}_I(m) - \left|h_x(m) - \bar{h}_I(m)\right|}{\Lambda I(m)/\sqrt{m}}\right) \tag{2}$$

where $H(\cdot)$ is the cumulative density function of $h_x(m)$. Similarly, the probability for x being a non-photograph is computed as follows:

$$p_P(x) = \frac{2}{M}\sum_{m=1}^{M} H\left(\frac{\bar{h}_P(m) - \left|hx(m) - \bar{h}_P(m)\right|}{\Lambda_P(m)/\sqrt{m}}\right) \tag{3}$$

The decision is made by comparing the test results from the above two *z*-tests:

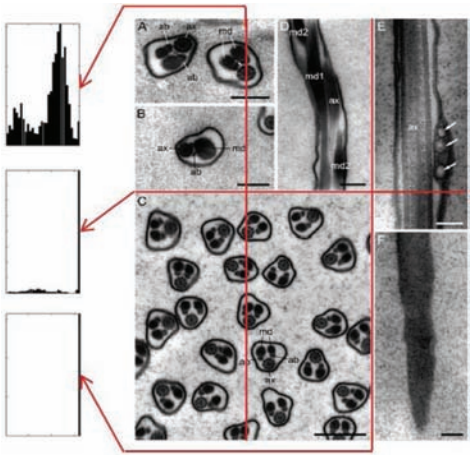
$$x = \begin{cases} \text{Photograph} & p_I(x) \gg p_P(x) \\ \text{Non-photograph} & p_I(x) \ll p_P(x) \\ \text{Combination} & \text{Otherwise} \end{cases}. \tag{4}$$

3.2 Panel segmentation

Panels in a figure are usually separated with margins in dark or light colours. Besides colour variations, the width and length of these margins are inconsistent between figures or within a figure, which prevents a universal colour-based threshold to be general applicable. As shown in Figure 1(b), a vertical margin splits the entire figure into two parts and within each part, more panels can be found.

For ease of reference, we name the margins that strike through a figure (or subfigure) as long margin (an example is shown in Figure 4) and the others as short margins. The long margins are relevant to the figure (or subfigure). That is, for any long margin 0, there exists one subfigure x^1 such that θ strikes through x^1 and separates x^1 into two complete halves. A short margin cannot split a figure (or subfigure) into two parts without having to disrupt the image integrity. An example of short margin is the boundaries of a zoom-in view.

Figure 4 Histograms for long margin detection. Histograms at the highlighted (in red) columns and row are shown. The histogram of the long margin depicts a spike at one end of the gray scale; whereas the others show spread-out distribution of pixel counts (see online version for colours)



There are two components in our panel segmentation process:

Half panel extraction: Figure is evaluated and parsed into two halves using long margins until each subfigure consists of one photograph.

Enclosed panel extraction: Zoom-in views are detected and the enlarged view is extracted from the subfigure.

This component involves detection of short margins and extraction of enclosed panel.

To detect the long margins in a figure, we generalise the idea described in Murphy et al. (2001) and analyse the entire figure for horizontal and vertical arrays of pixels that are homogeneous in colour and run across the figure. A histogram, denoted with h, is generated for each column (or row) of pixels. Without loss of generality, our discussion in the rest of this section is based on margins in white colour. Ideally, a long margin forms a single spike at the end of the histogram range. The height of this spike is equivalent to the height or the width of the figure. That is, for an m by N 8-bit gray scale figure x, the vertical long margin has the peak at 255 and the height of this peak is equal to M.

However, the ideal case seldom exists in published figures due to quantisation errors or noise. An example is shown in Figure 4. The histogram of a long margin is depicted on the bottom left of Figure 4. Small values spread into bins other than 255. The middle left histogram shows a row of pixels that consists of a long margin with part of a subfigure. The top left histogram shows the row of pixels that are part of three subfigures. To overcome colour variation of the pixels along the margin, a tolerance parameter, denoted as t, is used in the segmentation process. Hence, the new histogram, h^*, becomes the integration of a square window with a width of t and the original histogram h. Multiple adjacent rows or columns are merged and the centreline is used to split the figure into two parts. Each part is then analysed for long margins recursively, until no more long margins can be identified.

Short margins mark the boundaries of a highlighted view in figures. To detect short margins, we employ the Hough transform to find horizontal and vertical edges. Figure 5(c) illustrates the Hough transform of a panel with a zoom-in view.

Two heuristics are used in our detection method:

1 only straight lines at 0 and 90 degrees are considered

2 the straight lines are located within the width and height of the panel.

As shown in Figure 5, at 0 degree, three significant peaks are identified. The two extreme peaks correspond to the top and bottom boundaries of the rectangular subfigure. The locations of these two points can be easily decided, since their coordinates are determined by the height of the panel. The third peak marks one horizontal short margin of the subfigure. Similarly, by examining the 90 degree and applying the constraint of panel width, we can find a vertical line that forms the other short margin of the subfigure.

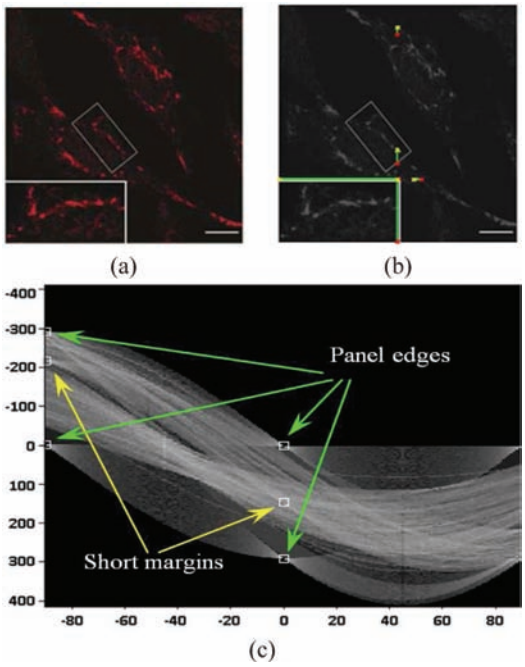
The reconstruction from peaks in the Hough space reveals the lines in the image space. Ideally, the line segments outline the short margins of the subfigure. However, it is common to have small line segments spreading across the image that are coinciding with the short margins. Image morphology is used to connect possible disconnected lines and small segments are then removed from the reconstruction. An example is shown in Figures 5(a) and (b). The detection result is highlighted with green in Figure 5(b) and the start and end points of a line segment are denoted with yellow and red dots.

4 Experimental results and discussion

4.1 Data preparation

Given a vast variety of biological and medical papers, we roughly categorised the figures into three types: figures that contain only non-photographs, figures that contain only photographs

Figure 5 Histograms for long margin detection. Histograms at the highlighted (in red) columns and row are shown. The histogram of the long margin depicts a peak at one end of the gray levels; whereas the others show scattered histograms (see online version for colours)



and figures that contain both. All figures were extracted from papers in PDF format and saved into JPEG images. Figures 1, 6(a), 7(a) and 8(a) illustrate sample instances used in our experiments. These figures are representative in that they are from papers in a variety of journals. As shown in these examples, figures in biological and medical papers are usually organised in a less regular manner and typically mix plots and photographs to highlight the findings with exemplar images.

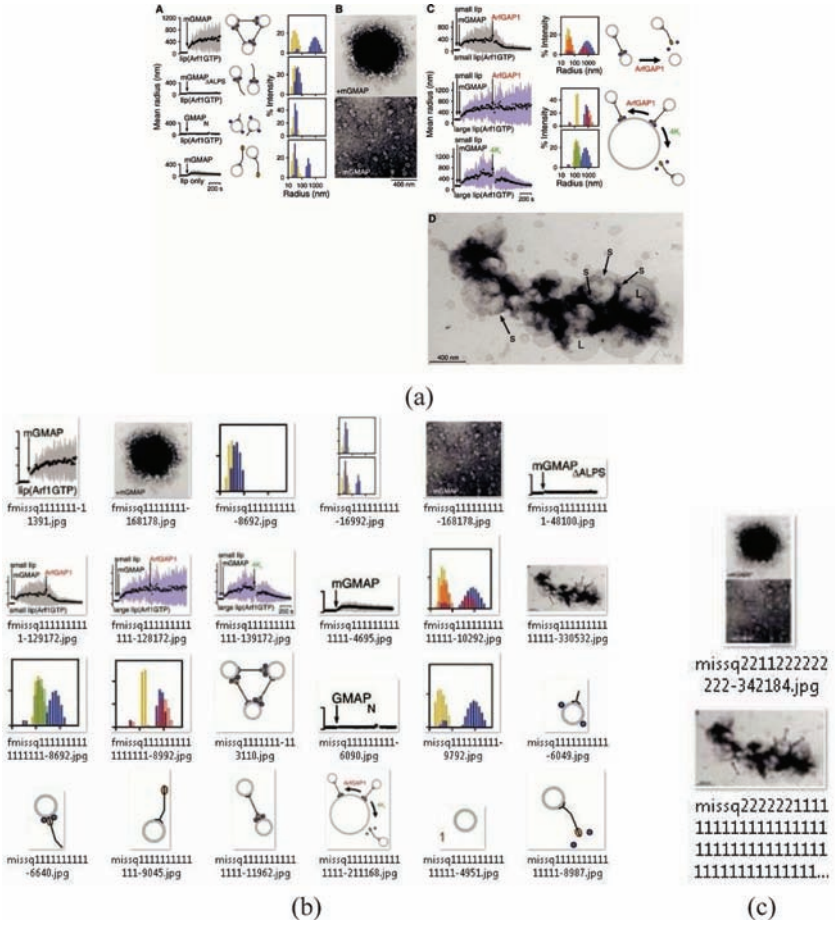
Two sets of research papers are retrieved separately and independently from PubMed Central using different keywords. One set contains 25 papers and the other contains 24 papers. There is no duplication in our collection. Experiments with the images from these two sets of papers were conducted independently.

From the first 25 papers, 182 figures were automatically extracted; from the other 24 papers, 178 figures were extracted. Figure captions were removed during the extraction process, whereas annotations and subfigure indexes were retained because they were integrated into the figures. All figures are converted into 8-bit gray scale images for processing. Table 1 summarises data sets used in our experiments.

4.2 Method implementation

Our methods were implemented with MATLAB. In the Hough transform used to detect highlighted views, we used the following parameters: angle resolution 1°, angle range [–90°, +90°] and translation step-size 1 pixel. These parameters are intuitive and generic, as demonstrated in our experimental results and can be successful in a variety of figures.

Figure 6 Experimental results: (a) original figure; (b) results produced with our method and (c) results produced with the benchmark method (see online version for colours)



For comparison, we used the method reported in Murphy et al. (2001) as the benchmark method. Following the description in the paper, we implemented it with MATLAB as well. It is our observation that choice of threshold affects the method performance significantly. We used an empirically selected gray scale threshold at 235, which produced the best segmentation results based on human verification.

Three experimental results are shown in Figures 6–8. The extracted images are shown as icons with a file name displayed underneath each icon. Figure 6 illustrates an example of a collage of photographs, plots and diagrams. Our method was successful in extracting all photographs and plots (see Figure 6(b)). Due to the presence of clear separating margins between components of a diagram, some diagrams are divided into small pieces. The results from the benchmark method extracted photographs, only one of which is further dividable (see Figure 6(c)). As stated in Murphy et al. (2001), non-photographs are not their focus; hence, it is understandable that no plots or diagrams are extracted using the benchmark method, which is also demonstrated in Figure 8, In which figure only plots exist.

Figure 7 Experimental results: (a) original figure; (b) results produced with our method and (c) results produced with the benchmark method (see online version for colours)

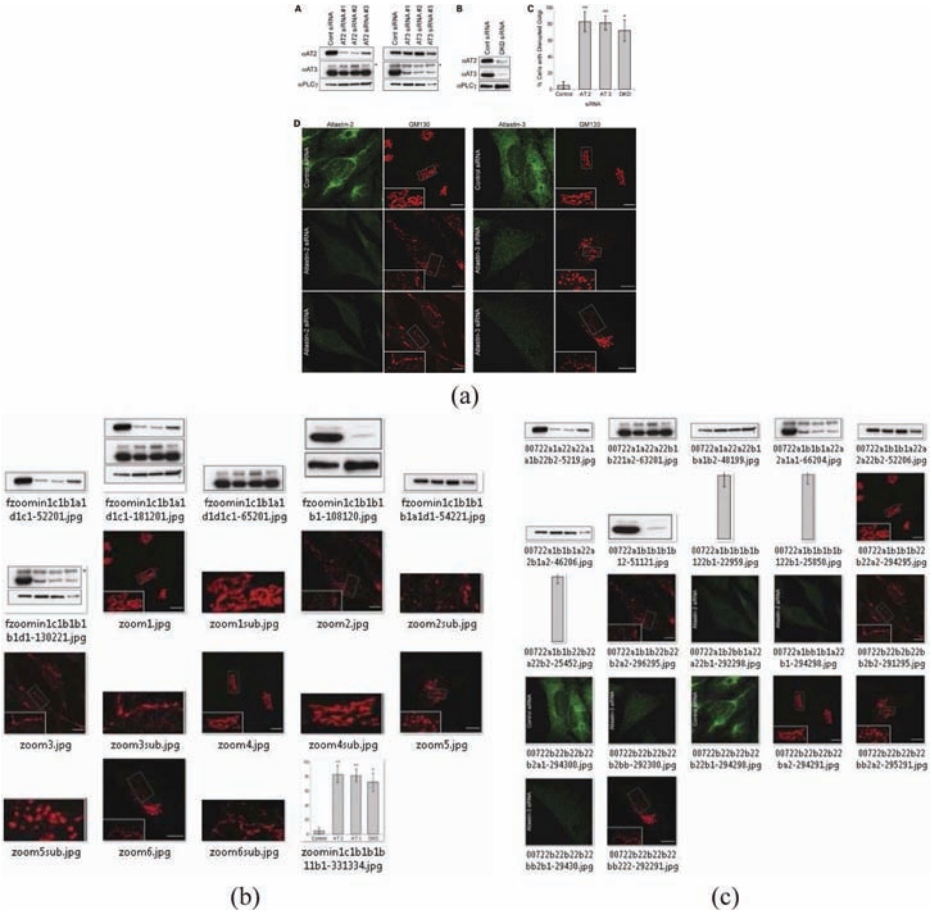


Figure 8 Experimental results: (a) original figure and (b) results produced with our method. The benchmark method was not successful in segmenting the figure in (a). The extracted panels are many small fragments

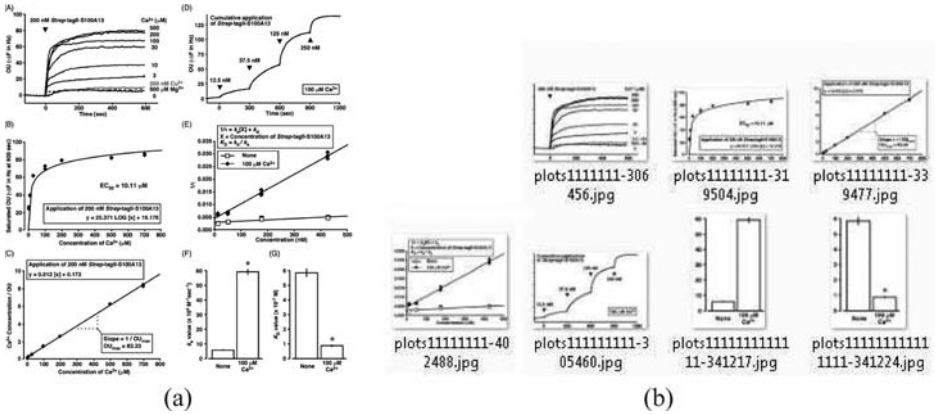


Table 1 Data sets used in our experiments

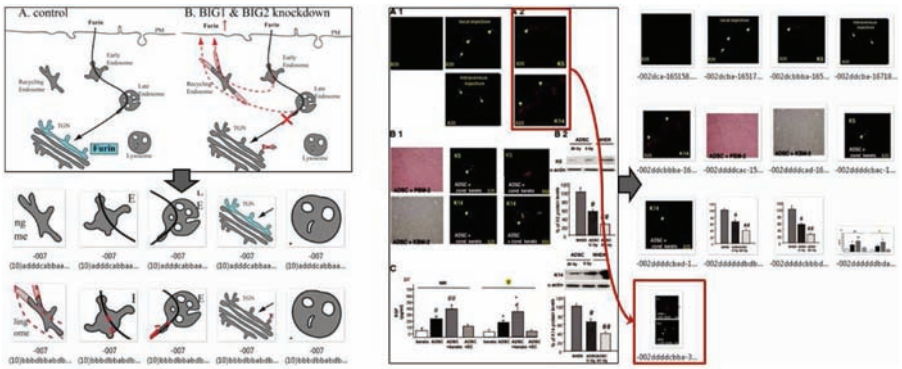
			Figure extracted	
Set	*# of papers*	*Photographs*	*Combination non-photographs*	
1	25	99	45	38
2	24	82	41	55

Figure 7 illustrates an example with highlighted views. The highlighted views are superimposed on their original photographs. Both methods successfully extracted photographs from the figure. However, it is clear that our method identified the highlighted views embedded in the figures and saved them into files. It is interesting to notice that the bar plot was extracted as a unity with our method, but was treated as photographs using the benchmark method and saved into three individual files.

Figure 8 illustrates an example that contains only non-photograph panels. Our method extracted all panels and saved them into individual files. The two plots on the bottom right were divided into two files.

In our experiments, two common segmentation errors usually occur, including over segmentation and under segmentation. In the case of over segmentation, figures are divided into more pieces than necessary. An example is shown in Figure 9(a). In the case of under segmentation, the processing stops before meaningful panels are properly extracted. In other words, figures can be further divided. An example is shown in Figure 9(b).

Figure 9 Examples of segmentation errors. The original images and the segmentation results are shown. The arrows point from the original image to the results. (a) shows an over segmentation example and (b) shows an under segmentation example, in which panel *B*1 (highlighted with a box) should be further divided into two (see online version for colours)



Tables 2 and 3 present the results using our method and the benchmark method. Experimental figures are categorised into photographs, non-photographs and combinations (i.e. collages of photographs and non-photographs). The results from the two groups of independent experiments agree with minimum difference. The correct rate listed in the last column of both tables is a ratio of the correctly segmented figure count vs. the total number of figures.

The correct rate of our method is close to 90% or above whereas the rate of the benchmark method is about 60%. The overall improvement rate of our method is 49.4%.

Table 2 Results from data set 1 of 182 figures. Cor. stands for correctly segmented; Ove. stands for over segmentation and Und. stands for under segmentation

	Photographs			*Combination*			*Non-photographs*			
Methods	*Cor.*	*Ove.*	*Und.*	*Cor.*	*Ove.*	*Und.*	*Cor.*	*Ove.*	*Und.*	*Cor. Rate*
Our	97	0	2	39	1	5	31	7	0	91.8%
Benchmark	97	0	2	10	35	0	9	29	0	63.7%

Table 3 Results from data set 2 of 178 figures. Cor. stands for correctly segmented; Ove. stands for over segmentation and Und. stands for under segmentation

	Photographs			*Combination*			*Non-photographs*			
Methods	*Cor.*	*Ove.*	*Und.*	*Cor.*	*Ove.*	*Und.*	*Cor.*	*Ove.*	*Und.*	*Cor. rate*
Our	79	1	2	30	2	9	48	7	0	88.2%
Benchmark	81	0	1	15	26	0	5	50	0	56.7%

The categorical error rate is summarised in Table 4. Both methods achieved highly satisfactory performance in processing photographs. It is clear that the benchmark method tends to over segment non-photographs as well as combinations and the error rates of these two cases are at 84.9% and 70.9%, respectively. In contrast, our method demonstrated greater robustness in dealing with combinations and non-photographs. The error rates for all cases are under 20% and the error rate for photographs is less than 3%.

Table 4 Overall error rates. Cor. stands for correctly segmented; Ove. stands for over segmentation and Und. stands for under segmentation. The values are in percentage

	Photographs			*Combination*			*Non-photographs*		
Methods	*Ove.*	*Und.*	*Total*	*Ove.*	*Und.*	*Total*	*Ove.*	*Und.*	*Total*
Our	0.5	2.2	2.7	3.5	16.3	19.8	15.1	0	15.1
Benchmark	0	1.7	1.7	70.9	0	70.9	84.9	0	84.9

Out of 360 figures in our data sets, there are 91 panels that contain highlighted views, of which 80 were successfully processed and highlighted views are separated from the panel. The error rate for processing highlighted views is 12.1%. As observed in our experimental results, the error of our method was mostly caused by multiple highlighted views. The benchmark method does not take this case into consideration and, hence, extracted no highlighted views.

Experiments were conducted on a desktop computer with Intel Core 2 Duo 3GHz, 4GB memory and running a 64-bit Windows 7 operating system. It took an average of 6.7 seconds and 0.5 seconds to process one figure using our method and the benchmark method, respectively. Although our method is slower than the benchmark method, its speed is satisfactory even for online applications.

5 Conclusion

In this paper, we present a model-driven, hierarchical method to extract panels from figures embedded in scientific papers. The results of our method will serve as a valuable component in automatic document image understanding and text and image-based full-text literature mining. Figures in papers are usually collages of multiple panels of the same type or different types. In addition, highlighted views carry information that is sometimes beyond the scope of that paper. To analyse these figures and extract information, a key step is to split figures into panels such that each one contains only a single image for image understanding. The method presented provides a means for automatic figure classification and segmentation.

Our method consists of two integrated components: figure classification and panel segmentation. The figure classification decides if there exists a photograph in the figure or panel. The mixture of plots and photographs is divided into subfigures. The splitting process repeats until no further panel collage can be identified. In all photograph panels, detection of highlighted views is also achieved. Our method transforms a photograph into a Hough space with a line template. Using reconstruction from Hough peaks, enclosed panels are retrieved and saved into separate files.

Experiments were conducted with a total of 360 figures extracted from two sets of papers that are retrieved with difference sets of keywords. These figures represent the diversity of figures published in the biomedicine field. From our experiments, the following conclusions are drawn:

- Despite vast differences among figures, our method successfully segmented figures and extracted photographs and non-photographs. Results from the two groups of independent experiments agree with minimum difference.
- The accuracy of our method is greater than the state-of-the-art methods and is superior when dealing with non-photographs and combinations of photographs and non-photographs. The improvement rate against our benchmark method is 49.4%.
- Our method was able to identify and extract highlighted views, which, to our best knowledge, is not possible for any existing method. The error rate in dealing with extraction of highlighted views is 12.1%.
- The computational time of our method is competitive. The average single figure processing time is at 6.7 sec. This enables online implementation.
- In our future work, we plan to employ texture to achieve a better categorisation of panels and circumvent over segmentation by integrating information derived from figure captions.

References

Arias, J., Chhabra, A. and Misra, V. (1998) 'A practical application of graphics recognition: helping with the extraction of information from telephone company drawings', *Graphics Recognition Algorithms and Systems*, Vol. 1389, pp.314–321.

Arias, J., Lai, C., Chandran, S., Kasturi, R. and Chhabra, A. (1993) 'Interpretation of telephone system manhole drawings', *International Conference on Document Analysis and Recognition*, Tsukuba Science City, Japan, pp.365–368.

Blostein, D., Lank, E. and Zanibbi, R. (2000) 'Treatment of diagrams in document image analysis', *First International Conference on Theory and Application of Diagrams*, London, UK, pp.330–334.

Doermann, D. (1998) 'The indexing and retrieval of document images: a survey', *Computer Vision and Image Understanding*, Vol. 70, No. 3, pp.287–298.

Drin, G., Morello, V., Casella, J-F., Gounon, P. and Antonny, B. (2008) 'Asymmetric tethering of flat and curved lipid membranes by a golgin', *Science*, Vol. 320, No. 5876, pp.670–673.

Fisher, J.L. (1991) 'Logical structure descriptions of segmented document images', *International Conference on Document Analysis and Recognition*, Saint-Malo, France, pp.302–310.

Fisher, J.L., Hinds, S.C. and DAmato, D.P. (1990) 'A rulebased system for document image segmentation', In *the 10th ICPR*, Vol. 1, pp.567–572.

Lu anb X., Kataria, S., Brouwer, W.J., Wang, J.Z., Mitra, P. and Giles, C.L. (2009) 'Automated analysis of images in documents for intelligent document search', *International Journal on Document Analysis and Recognition*, Vol. 12, No. 2, pp.65–81.

Mao, S. and Rosenfeld, A. (2003) 'Document structure analysis algorithms: a literature survey', In *SPIE*, Vol. 5010, pp.197–207.

Murphy, R.F., Velliste, M., Yao, J. and Porreca, G. (2001) 'Searching online journals for fluorescence microscope images depicting protein subcellular location patterns', *2nd IEEE International Symposium on Bioinformatics and Bioengineering*, Washington, DC, pp.119–128.

Niyogi, D. and Srihari, S.N. (1995) 'Knowledge-based derivation of document logical structure', *International Conference on Document Analysis and Recognition*, Washington, DC, USA, pp.472–475.

Qian, Y. and Murphy, R.F. (2008) 'Improved recognition of figures containing fluorescence microscope images in online journal articles using graphical models', *Bioinformatics*, Vol. 24, No. 4, pp.569–576.

Samuel, J., Yuan, X., Yuan, X.J. and Walton, B. (2010) 'Mining online full-text literature for novel protein interaction discovery', *International Workshop on Data Mining for High Throughput Data from Genome-wide Association Studies*, Hong Kong, China, December 2010.

Shatkay, H., Chen, N. and Blostein, D. (2006) 'Integrating image data into biomedical text categorization', *Bioinformatics*, Vol. 22, pp.e446–e453.

Tsujimoto, S. and Asada, H. (1990) 'Understanding multi-articled documents', *International Conference on Pattern Recognition*, Atlantic City, NJ, USA, pp.551–556.

Xie, S. and Hua, B. (2010) 'Sperm ultrastructure in two species of panorpa and one bittacus', *Micron*, Vol. 41, No. 6, pp.622–632.

Yamashita, A., Amano, T., Takahashi, I. and Toyokawa, K. (1991) 'A model based layout understanding method for the document recognition system', *International Conference on Document Analysis and Recognition*, Saint-Malo, France, pp.130–138.

Zheng, Y., Li, H. and Doermann, D. (2005) 'A parallel line detection algorithm based on hmm decoding', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 5, pp.777–792.