CIS UDEL Working Notes on ImageCLEF 2015: Compound figure detection task

Xiaolong Wang, Xiangying Jiang, Abhishek Kolagunda, Hagit Shatkay* and Chandra Kambhamettu*

Department of Computer and Information Sciences,
University of Delaware, Newark, DE US
{xiaolong,jiangxy,abhi,shatkay,chandrak}@udel.edu

Abstract. Figures that are included in biomedical publications play an important role in understanding essential aspects of the paper. Much work over the past few years has focused on figure analysis and classification in biomedical documents. As many of the figures appearing in biomedical documents comprise multiple panels (subfigures), the first step in the analysis requires identification of compound figures and their segmentation into subfigures. There is a wide variety ways to detect compound figures. In this paper, we utilize only visual information to identify compound vs non-compound figures. We have tested the proposed approach on the ImageCLEF 2015 benchmark of 10,434 images; our approach has achieved an accuracy of 82.82%, thus demonstrating the best performance when compared to other systems that use only visual information for addressing the compound figure detection task.

Keywords: Compound figure detection, visual information, biomedical image analysis, image classification, ImageCLEF 2015.

1 Introduction

Figure classification and understanding within the biomedical literature has attracted much research interest over the past years. Figures included in biomedical publications form a necessary source of knowledge and understanding. Most of the works on image analysis within biomedical documents aim at recognizing different biomedical image categories. Notably, figures appearing within biomedical documents are often *compound*, that is, they comprise multiple panels that are typically referred to as *subfigures*. Categorization and analysis of images usually requires working at the subfigure level, and as such, a primary step in the analysis is the identification of compound figures and their segmentation into subfigures [2, 6].

Over the past few years, several approaches were proposed for compound figure segmentation, within the field of biomedical image retrieval [2, 8, 6]. Previous

* These authors contributed equally to this work.

work can be categorized into two main schemes: The first is based on the analysis of peak region detection within the image; the peak region is then used as a reference to find separating lines for segmentation [2, 8]. The main drawback of this scheme is that it is susceptible to noise and may lead to over-segmentation [8]. This issue is especially prevalent in irregular compound figures, where the separators between different subfigures do not cut across a complete row or column. Moreover, setting up the threshold value for segmentation with respect to a peak region is not straightforward — different thresholds usually lead to different results. For instance, Chhatkuli *et al.* [2] set the threshold at 0.97 times the maximum value in a given figure. This threshold value is based on manual tests over the training data. Another factor is the occurrence of text within figures. As text is irregular, it can be an obstacle for obtaining the segmentation lines [2]. Removing text from a compound figure usually plays an important role in the final result.

The second scheme is based on connected components analysis [5], as was done in earlier work [6]. The general idea is to evaluate the connectivity among different subfigures within a compound figure using visual information. Connected components analysis groups the pixels into different components using similarity in pixel values. Pixels in each resulting component share similar values. Once different connected regions are formed, the boundary between different regions can be used as segmentation lines separating different components. In this work, the analysis of connected components is applied first to the given figure, while we also add several post-processing steps. These post-processing steps help improve compound figure detection. We then integrate the two different schemes. The experimental results demonstrate that the fusion scheme can help improve performance compared to each of the individual schemes applied alone.

The rest of the paper is organized as following. In Section 2, we provide an overview of the datasets. The proposed compound figure detection approach is discussed in Section 3. The analysis of component connectivity based scheme is discussed first, followed by a presentation of the peak region detection scheme and the fusion scheme. Section 4 presents the experimental results submitted to the ImageCLEF 2015. In the end, conclusions are given in Section 5.

2 Dataset

In our experiments, we use the dataset provided in the ImageCLEFmed 2015 benchmark. We refer the reader for more details to the respective task description [7, 3, 1, 4]. In this report, we focus on the medical image classification task [3] and specifically on compound figure detection. Notably, we use only visual information for addressing this task.

3 Approach

In this report, we first discuss the proposed compound figure detection scheme, where we illustrate the details of our detection method, utilizing only visual in-

formation. As shown in the ImageCLEF15 comparison of the results with those obtained by other systems, our approach achieves the highest level of performance among schemes that use only visual information, while its accuracy is only 2.57% lower than that of the top performing scheme, which combines visual and textual information.

3.1 Connected Component Analysis Based Scheme

The first part of our compound figure detection scheme is based on the analysis of component connectivity of subfigures in an image [5]. This scheme is based on graph traversal theory. The general idea is to determine the connectivity of the current pixel to neighboring pixels based on pixel-intensity; the method is both effective and simple to implement.

The general scheme of connected component analysis used in our work is as follows: First, RGB images are converted into grayscale images. Then we rescale the pixel intensities in the whole image into values in the range [0, 1]. The underlying assumption is that the boundary between subfigures typically consists of white pixels. The white area is defined as the region where pixel values are consistently greater than 0.9; other regions are defined as black regions. By comparing the image intensities to this threshold value, we can get the mask image M. M is a binary image as indicated in Fig. 1. In this work, the white color represents the foreground and black indicates the background region. The connected components are extracted based on the mask binary image M.

After that, we scan the resulting, simplified image pixel-by-pixel (top to bottom and left to right). Connected regions in which adjacent pixels share a similar range of intensity values $[v_0, v_1]$ are identified. The connected components labeling operator scans the whole image by moving along each row until it reaches a pixel p which has not been previously labeled. If pixel p is not labeled in the previous stage, we examine two p's predecessor neighboring pixels directly up (denoted p_u), and to the left (p_l). The label value assigned to pixel p is based on the comparison with these two neighboring pixels.

After scanning the whole image, each detected component in the figure is labeled with a different value. An important issue for compound figure detection is to minimize the influence of false positive area where non-compound regions are misclassified as compound regions. Most of these false positive areas are caused by the connected text. To address this issue, rather than directly removing text from the images [2], we apply a criterion based on the ratio evaluation among regions' areas. Two different ratio criteria are used in this work. The first ratio value T_{r1} is defined as the ratio between area of the detected subfigure and the whole figure. If T_{r1} is smaller than 0.1, then this region is classified as false positive. The second ratio value T_{r2} is calculated based on the area ratio between the detected components and the maximum component. If the ratio value T_{r2} is smaller than 0.15, the detected region is classified as false positive. This setting has proven effective in our experiments as illustrated in Fig. 1.

The illustration of the whole scheme is presented in Fig. 4. If more than one subfigure is detected, the given figure is classified as compound figure, otherwise

not. To handle compound figures separated by black rather than white regions, we invert all images and perform subfigure detection using the same procedure.

3.2 Peak Region Detection Based Scheme

Besides the connected component analysis method described above, we also test the performance of directly using pixel intensity to segment the figure as used in works [2, 8]. The idea of this method is to find white margins based on the pixel intensity. As indicated in Fig. 2, the images are scanned in two directions, namely along the x-axis and along the y-axis. Both scanning processes are conducted iteratively until no more white margins are detected.

Consider an image I represented as a matrix $I(x, y)$, where x is the row index and y is the column index; let W and H be the total number of rows and columns, respectively. Assuming that the subfigures are separated by white margins, the first step is pixel projections operation along the x-axis and along the y-axis as indicated in Fig. 2. Formally:

$$\begin{aligned} I_x &= min\, I(x, y) \text{ y}\in [1, ..., H], \\ I_y &= min\, I(x, y) \text{ x}\in [1, ..., W], \end{aligned} \tag{1}$$

that is, I_x is a candidate separating row and I_y is a candidate separating column. The next step is to find a peak region within I_x and I_y. The peak region indicates an area located within the continuous region whose pixel value is greater than a predefined threshold. In this work, considering the noise and other influential factors, the threshold is set to 0.85 times of the maximum pixel intensity in the whole image. By comparing with the threshold, we can find the peak region along I_x and I_y vector. From this, we obtain the index and the region width. These peak regions are regarded as the margin between subfigures. Based on these detected margins, the subfigure region is then calculated.

For a specific testing image, to get rid of false positives and minimize the influence of the text region, two different post-processing steps are applied. First, we set a threshold on the minimum area of a detected peak region. Another criterion is to measure the ratio value calculated between the current segmented area and the maximum segment detected. If the ratio is smaller than 0.3, the detected segmentation region is classified as false positive. If more than one sub-region is detected, input figure is classified as compound, otherwise not. As before, this method assumes that the separation between sub-figures consists of white pixels. To also consider black separators, if the figure is not classified as compound, we invert the image and go through the processing steps discussed above again.

3.3 Fusion Scheme

In our work, we also fuse the above two different schemes – connected component analysis is used as the first step and if no compound figure is detected, peak region detection is applied as the second step.

An illustration of the proposed scheme is shown in Fig. 3. Our results showed that connectivity component analysis is good at removing false positives caused by text regions. However, it is not as effective for detecting compound figures consisting of graph images (e.g. line graphs or diagrams). These types of compound figures can be detected using peak region detection approach. We conducted a standalone comparison between the proposed different schemes to evaluate their respective performance.

4 Experimental Results

We evaluated the proposed approach on ImageCLEF 2015 benchmark, which includes 10,434 different figures. Several illustrations of the experimental results are provided in Fig. 4. The overall accuracy is calculated as $accuracy = \frac{C_g}{C} \times 100\%$, where C_g represents the number of correctly detected figures and C is the total number of samples in the set. In addition, we also consider the well-known *recall* and *precision* measures, as shown in Table 1. The latter two measures are calculated as:

$$\begin{aligned} Precision &= \frac{TP}{TP + FP}, \\ Recall &= \frac{TP}{TP + FN}, \end{aligned} \tag{2}$$

where TP is the number of true positives (compound figures) detected by the proposed scheme, FP is the number of false positives (figures that are non-compound, but labeled as compound by our scheme), and FN is the number of false negatives (figures that are compound, but not detected as such by the proposed approach).

As listed in Table 1, the connected component analysis based scheme performs better than the peak-region detection based scheme. By combining the two different schemes, we have obtained an accuracy of 82.82% on the test dataset.

For the sake of completeness, we also demonstrate several cases, shown in Fig. 5, in which our system fails to detect or to correctly segment a compound figure. As illustrated by Fig. 5(a), when the boundaries between subfigures are thin, although our algorithm can correctly classify the given compound figure, the sub-figure segmentation does not work well. Moreover, segmenting diagrams remains a challenge, as indicated in Fig. 5(b). Over-segmentation is still a common problem for this kind of non-compound figures [8].

5 Conclusion and Future Work

In this work, we have studied the problem of compound figure detection. Two different schemes, as well as an integration of the two, are evaluated. Our integrated scheme outperforms the other systems that use only visual information,

Table 1. Comparison results between proposed approaches.

Approach	Accuracy	Precision	Recall
Component connectivity analysis	82.47%	82.48%	72.84%
Peak region detection	81.04%	84.94%	73.23%
Fusion scheme	82.82%	86.06%	69.49%

participating in this challenge, by more than 10%. In this challenge, the only system outperforming this system (by 2.57%) used a combination of textual and visual information.

6 Acknowledgment

This work was supported by NIH Award 1R56LM011354-01A1.

References

1. M. A. Amin and M. K. Mohammed. Overview of the ImageCLEF 2015 medical clustering task. In *CLEF2015 Working Notes*, CEUR Workshop Proceedings, Toulouse, France, September 8-11 2015. CEUR-WS.org.
2. A. Chhatkuli, A. Foncubierta-Rodríguez, D. Markonis, F. Meriaudeau, and H. Müller. Separating compound figures in journal articles to allow for subfigure classification. In *SPIE medical imaging*, pages 86740J–86740J. International Society for Optics and Photonics, 2013.
3. A. García Seco de Herrera, H. Müller, and S. Bromuri. Overview of the ImageCLEF 2015 medical classification task. In *Working Notes of CLEF 2015 (Cross Language Evaluation Forum)*, CEUR Workshop Proceedings. CEUR-WS.org, September 2015.
4. A. Gilbert, L. Piras, J. Wang, F. Yan, E. Dellandrea, R. Gaizauskas, M. Villegas, and K. Mikolajczyk. Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. In *CLEF2015 Working Notes*, CEUR Workshop Proceedings, Toulouse, France, September 8-11 2015. CEUR-WS.org.
5. R. Gonzalez and E. Richard. *Digital Image Processing*. Prentice-Hall, 2002.
6. H. Shatkay, N. Chen, and D. Blostein. Integrating image data into biomedical text categorization. *Bioinformatics*, 22(14):e446–e453, 2006.
7. M. Villegas, H. Müller, A. Gilbert, L. Piras, J. Wang, K. Mikolajczyk, A. G. S. de Herrera, S. Bromuri, M. A. Amin, M. K. Mohammed, B. Acar, S. Uskudarli, N. B. Marvasti, J. F. Aldana, and M. del Mar Roldán García. General Overview of ImageCLEF at the CLEF 2015 Labs. Lecture Notes in Computer Science. Springer International Publishing, 2015.
8. X. Yuan and D. Ang. A novel figure panel classification and extraction method for document image understanding. *International journal of data mining and bioinformatics*, 9(1):22–36, 2014.

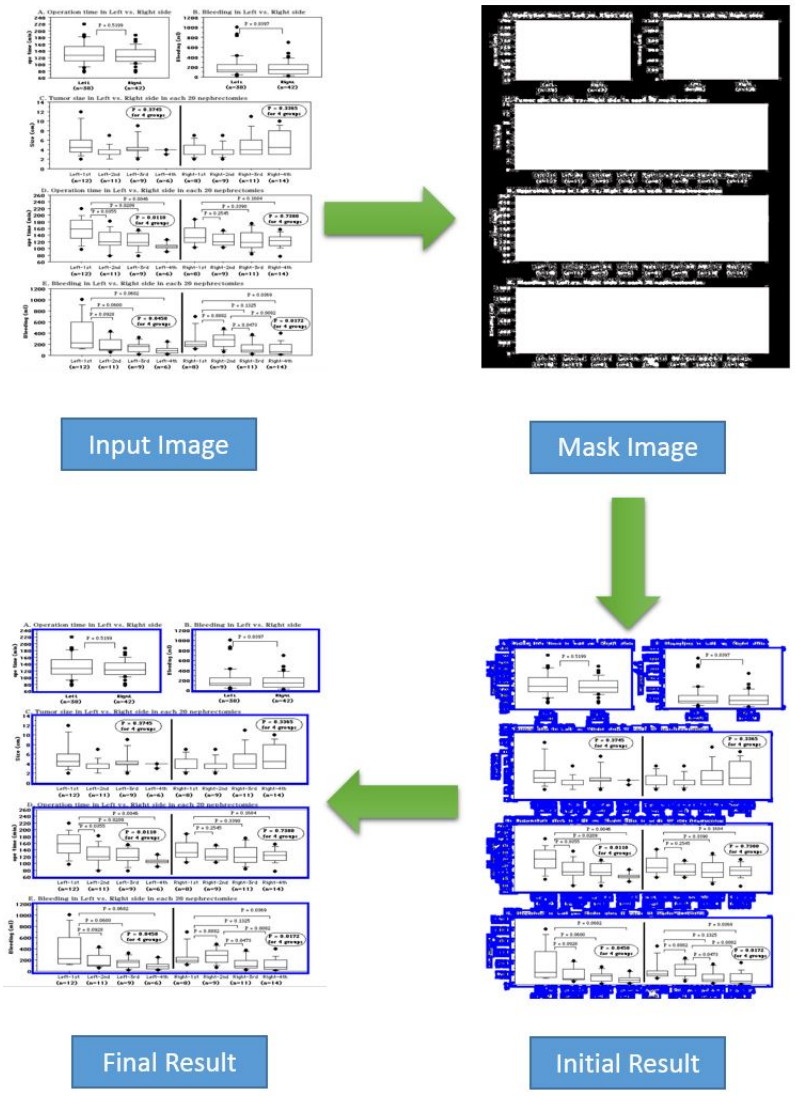


Fig. 1. Illustration of compound figure detection scheme using connected components analysis. The mask image correspond to M. The final result is based on the postprocessing of the initial figure result. The areas surrounded by blue lines are segmented regions.

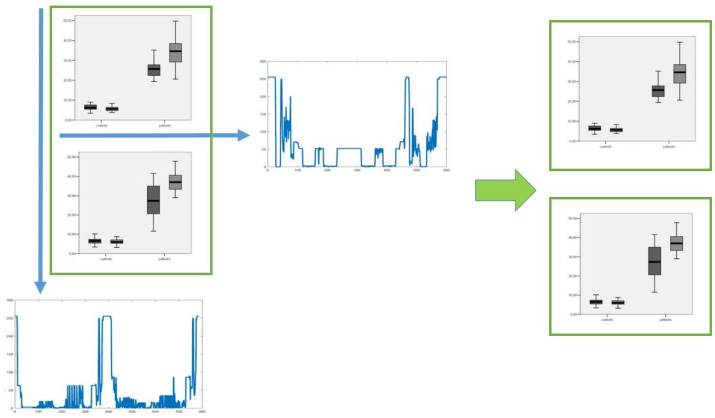


Fig. 2. Illustration of peak region detection based scheme. The minimum value of the projection along x-axis and y-axis is first obtained as indicated by the arrow. The continuous peak region calculated from the minimum value vector represents the long margin. Based on the peak region of of this vector, a compound figure is detected and segmented.

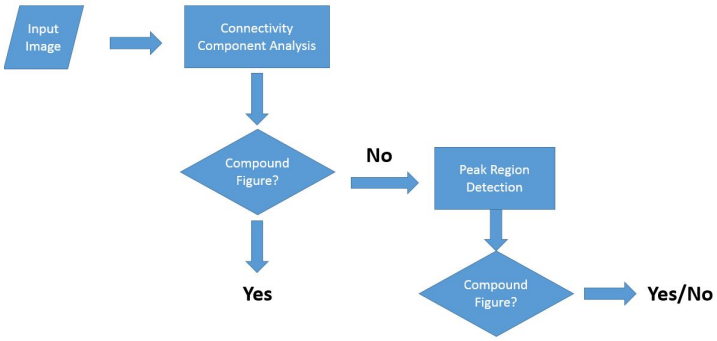


Fig. 3. Illustration of the proposed fusion scheme for compound figure detection.

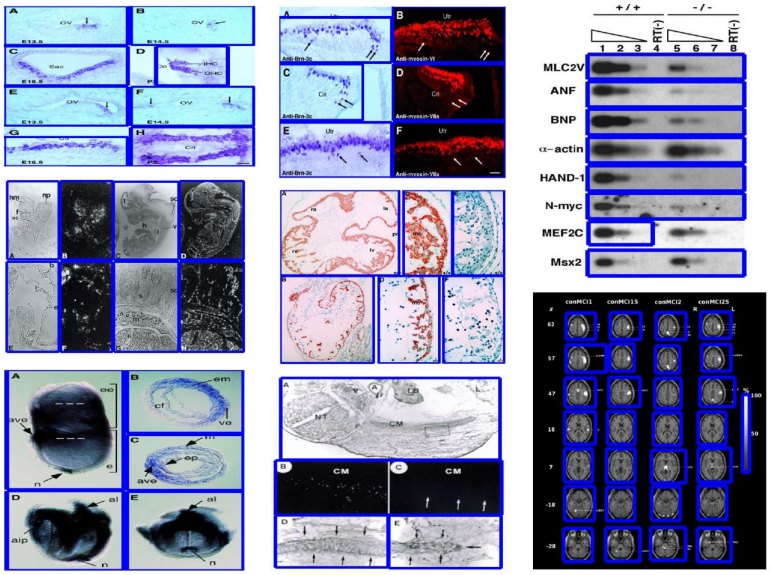
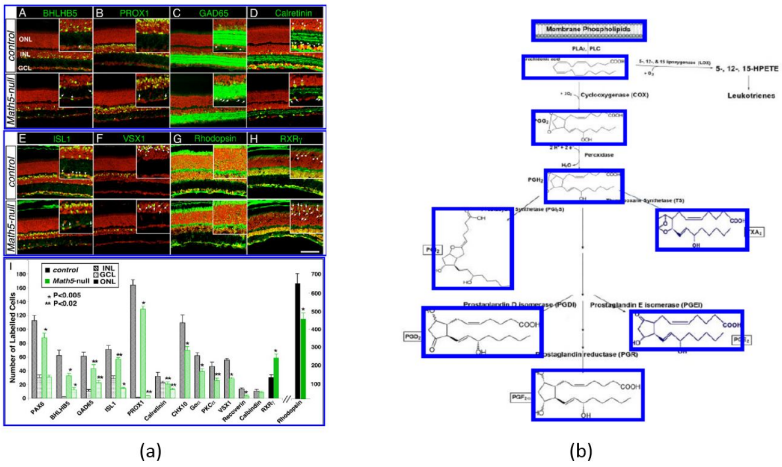


Fig. 4. Examples of several different compound figure segmentation results.



(a) (b)

Fig. 5. Illustration of two failure cases obtained in the experiments. (a) Under-segmentation of the compound figure. (b) Over-segmentation of the non-compound figure.