NLM at ImageCLEF 2015: Biomedical Multipanel Figure Separation

K.C. Santosh, Zhiyun Xue, Sameer Antani, and George Thoma

U.S. National Library of Medicine
National Institutes of Health, Bethesda, MD 20894.

Email. santosh.kc@nih.gov, xuez@mail.nih.gov, sameer.antani@nih.gov and george.thoma@nih.gov

Abstract. This paper summarizes the participation of the National Library of Medicine (NLM) in the imageCLEF 2015 biomedical multipanel figure separation task. In this task, our method uses two different techniques that are employed on the basis of characteristics of the figures: 1) stitched multipanel figure separation; and 2) multipanel figure separation with homogeneous gaps. Fusion of the two techniques achieved an accuracy of 84.64%.

Keywords: Biomedical articles, multipanel figure separation, content-based image retrieval.

1 Motivation

Medical image retrieval has been considered as an important research domain over the past 20 years [1, 2, 6, 15, 17, 21, 22]. Figures in the biomedical publications are often composed of multiple panels. Multipanel figures are used as an aid for grouping related visual artefacts for human consumption. However, they may comprise of images from different modalities (such as x-ray, MRI, CT, microscopy, graphics). In [4, 13], authors report an increasing use of visual material in biomedical publications. The average number of figures per article in the reputed biomedical journals ranges from 6 to 31 [7, 23]. More importantly, according to [11,12,16], multipanel figures represent about 50% of the figures in the biomedical open access image data sets such as those used in the imageCLEF (URL: http://www.imageclef.org) benchmark. Mixed modality in multipanel figures pose a challenge for image retrieval [1, 2, 15, 17] and modality classification systems [8,19,21]. We also note that these figures are not commonly available in biomedical publication datasets as standalone entities that could be readily used by automated systems since rarely do publishers require authors to submit figures (and captions) in separate files for easy access. In other words, most of the figures packaged as a single image file in the article thereby adversely affecting their accessibility by automatic multimodal indexing systems such as the National Library of Medicine's OPENi system (URL: http://openi.nlm.nih.gov) [14]. In this

context, multipanel figure separation is considered as a crucial step for high quality content-based image retrieval (CBIR) [3,5,6,14]. Therefore, we call this step 'a precursor' to biomedical CBIR.

The remainder of the article is organized as follows. Our method is explained in Section 2, where we provide details on two different panel-splitting techniques and their fusion. In Section 3, we present testing results and analysis. Finally, we summarize the paper in Section 4.

2 Methods

2.1 Outline

Uniform-space-separated multipanel figures comprise a significant subset of the imageCLEF benchmark data. These include regular (images) and graphical (illustrations, charts, plots) type figures. Pixel intensity profile-based and homogeneity-based (for crossing bands) methods are commonly used (and often sufficient) to separate the panels [3,5,5,18]. Other methods uses optical character recognition (OCR) for stitched or fully connected multipanel figures [3,14]. But, their solution is sensitive to common errors generated by the OCR and are rigid about the alignment of subfigure panel labels relative to each other. To the best of our knowledge, no methods have been reported that separate stitched multipanel figures purely from an image analysis standpoint. A primary challenge for image analysis-based techniques is that no clear boundaries and homogeneous gaps exist between fully connected panels. In this imageCLEF 2015 participation [10, 22], we combine two different techniques, operating separately to separate both stitched multipanel figures and the multipanel figures with homogeneous gaps. As a preliminary step, we overlook automating figure type selection (fully-connected and with homogeneous gaps), and focus on developing automatic techniques for separating the panels. Automatically detecting the figure types is left for future work. We manually separated the two types of multi-panel figures in the data set (see Fig. 1). Fig. 2 shows an example of stitched multipanel figure and two examples having homogeneous gaps between the panels.

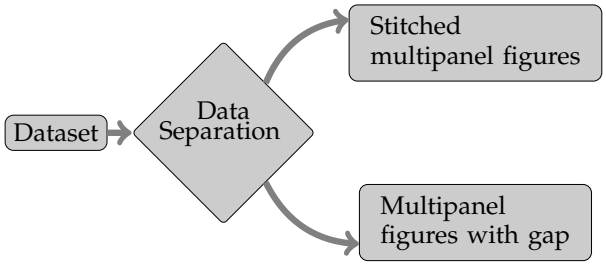
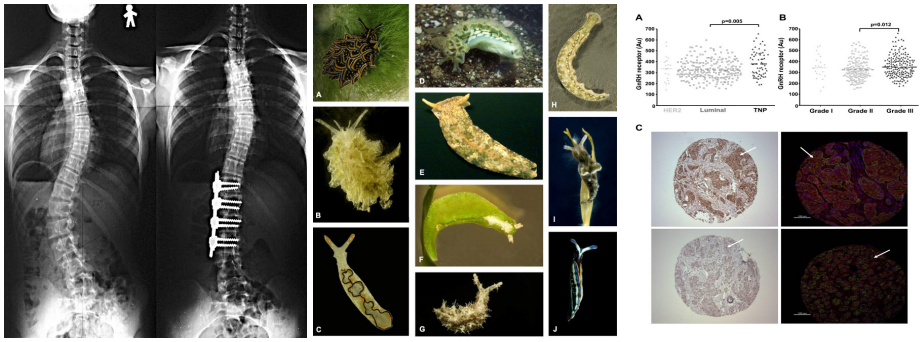


Fig. 1. Stitched (or fully connected) multipanel figures are manually separated from those with regular or homogeneous gaps between the panels.



(a) (b) (c)

Fig. 2. Both samples: (a) stitched multipanel figure, and (b) and (c) multipanel figures with homogeneous gaps (or crossing bands), are shown.

2.2 Stitched multipanel figure separation

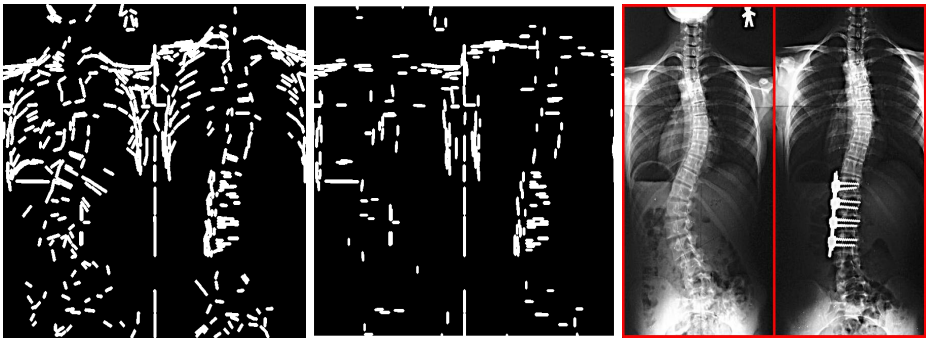
For stitched (i.e., fully connected) multipanel figures, we apply our previously reported technique [20]. The steps for stitched multi-panel figure separation can be summarized in the following two steps:

1) Line segment detection, and,
2) Line vectorization.

Details on this technique can be found in [20]. For completeness, we summarize the major steps below.

Line segment detection. The line segment detector (LSD) is designed to detect local straight contours (i.e., line segments), from the zones where the grey level changes from dark to light or vice-versa [9]. It uses edge pixel gradients to detect level lines for separating stitched panels. Fig. 3 shows an output of line segment detection.

Line vectorization. This step connects all prominent broken lines along the panel boundaries while eliminating unwanted line segments within the panels. Like other state-of-the-art techniques, it uses profile-based concept to connect lines from end to end (horizontal and vertical). Projection profiles from a 2D image $\mathbf{f}(x, y)$ of size $m \times n$ can be computed as $p_{\theta=\pi/2} = \sum_{1 \leq x \leq m} \mathbf{f}(x, y)$ and $p_{\theta=0} = \sum_{1 \leq y \leq n} \mathbf{f}(x, y)$. To eliminate dominant line segments that are typically resulted from the objects within the panels, we compute their corresponding profile transform (i.e., p_θ^2), which is then normalized by using their mean and standard deviation. As a consequence, the magnitude of the line segments along panel boundaries are more pronounced. To make it efficient, line segments are first filtered in two orthogonal directions: 0 and $\frac{\pi}{2}$, as shown in Fig.3.



(a) Line segments (b) Filtered line segments (c) Output (in red)

Fig. 3. An example showing (a) line segment detection, (b) Filtered line segments and (c) output: panels using rectangular boxes (in red).

2.3 Multipanel figure separation with homogeneous gaps

Since majority of the multipanel figures in the ImageCLEF 2015 dataset are separated by homogeneous horizontal or vertical crossing bands of uniform color, we apply our previously reported method [3]. It contains five distinct modules:

1) Text label extraction,
2) Panel subcaption extraction,
3) Panel segmentation,
4) Panel label extraction, and,
5) Combination of all outputs from the previous modules.

Note that in this participation, considering the dataset, the first two modules are not included since no figure caption text is provided.

Panel segmentation. The aim of this module is to identify homogeneous gaps (or crossing bands) for separating panels along them. Specifically, it is composed of five major steps: 1) image overlay/markup removal; 2) homogenous crossing band extraction; 3) border band (homogenous band that is located on the boundary of the panel) identification; 4) low gradient band (a band that does not have a sharp boundary line) removal; and 5) image division based on crossing bands. For images where the homogeneous gaps do not cross end-to-end, two iterations are required. For example, in Fig. 2 (b), the first iteration (that goes vertically) results three panels, which are still multipanel figures.

Panel label extraction. This module is designed to detect panel labels from each individual panel. It comprises of three steps: 1) panel label segmentation connected components (CCs); 2) CC recognition using OCR; and 3) refinement of OCR results to get panel labels. The module results several candidate sets of panel labels. In the combination module, the panel label candidate sets (obtained

Table 1. Performance comparison (multipanel separation rate in %). Runs are ranked based on the decreasing order of performance score.

Group name	Run type	Score
NLM run_2	Visual	**84.64**
NLM run_1	Visual	**79.85**
AAUITEC run_3	Visual	49.40
AAUITEC run_2	Visual	35.48
AAUITEC run_1	Visual	30.22

via panel label extraction) are matched with the panels (obtained via panel segmentation). The results of panel segmentation can help selecting the best label set while the results of panel label extraction can help splitting a panel further if multiple labels are found within it. For more detailed description, we refer readers to our previous work [3].

3 Experiments

3.1 Dataset and evaluation protocol

The imageCLEF 2015 panel segmentation dataset comprises two parts: training and test, composed of 3403 and 3381 images, respectively. It is important to note that our method does not use the training set. It separates every single image independently from test set without training. From the test set, we manually selected 145 images in the category of stitched multipanel figures. For more details about datasets and evaluation protocol, we refer to [10].

3.2 Results: comparative study

Following the method described in Section 2, we have submitted two different runs (designated as run_1 and run_2). In both runs, stitched multipanel figure separation (see Section 2.2) is combined. As described in Section 2.3, in run_1, panel separation is used while in run_2, panel label extraction is integrated with panel separation.

Table 1 shows an overall performance evaluation of our system and a comparison with other participants. Our results are reported as 79.85% and 84.64%, for run_1 and run_2. Since the performance of the stitched multipanel figure separation remains the same in both runs, the performance difference of approximately 5% in run_2 is attributed to panel label extraction. Panel label extraction does not only help improving the panel separation, but can be used to link the panel with its relevant caption fragment. Out of the two runs, we have received a best multipanel separation rate of 84.64%.

4 Summary

We have participated in imageCLEF 2015 biomedical multipanel figure separation task. We have submitted our test results by combining two different techniques. Our first technique separated panels from stitched multipanel figures, which is motivated by the fact that no state-of-the-art techniques reported any solutions. Our second technique focused on other remaining multipanel figures that are having homogeneous gaps between the panels. Based on the evaluation protocol designed by the organizer [10], our test outperforms the other participants by more than 35%.

Both techniques perform automatically but, their fusion is not since we have manually separated the dataset for them. As next steps, we plan to automatically categorize multipanel figures based on their characteristics into stitched multipanel and multipanel figures having homogeneous gaps.

Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC). The authors would like to thank Dr. Daekeun You (currently at the University of Michigan Health System) for his prior contributions that are part of the method used.

References

1. Aigrain, P., Zhang, H., Petkovic, D.: Content-based representation and retrieval of visual media: A state-of-the-art review. Multimedia Tools and Applications 3(3), 179–202 (1996)
2. Akgül, C.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B.: Content-based image retrieval in radiology: Current status and future directions. J. Digital Imaging 24(2), 208–222 (2011)
3. Apostolova, E., You, D., Xue, Z., Antani, S., Demner-Fushman, D., Thoma, G.R.: Image retrieval from scientific publications: Text and image content processing to separate multipanel figures. Journal of the American Society for Information Science and Technology 64(5), 893–908 (2013)
4. Aucar, J.A., Fernandez, L., Wagner-Mann, C.: If a picture is worth a thousand words, what is a trauma computerized tomography panel worth? The American Journal of Surgery 6(194), 734–740 (2007)
5. Cheng, B., Antani, S., Stanley, R.J., Thoma, G.R.: Automatic segmentation of subfigure image panels for multimodal biomedical document retrieval. In: Agam, G., Viard-Gaudin, C. (eds.) Document Recognition and Retrieval XVIII - DRR 2011, 18th Document Recognition and Retrieval Conference, part of the IS&T-SPIE Electronic Imaging Symposium, San Jose, CA, USA, January 24-29, 2011, Proceedings. SPIE Proceedings, vol. 7874, pp. 1–10 (2011)
6. Chhatkuli, A., Markonis, D., Foncubierta-Rodríguez, A., Meriaudeau, F., Müller, H.: Separating compound figures in journal articles to allow for subfigure classification. In: SPIE, Medical Imaging (2013)

7. Cooper, M.S., Sommers-Herivel, G., Poage, C.T., McCarthy, M.B., Crawford, B.D., Phillips, C.: The zebrafish {DVD} exchange project: A bioinformatics initiative 77, 439 – 457 (2004)
8. Demner-Fushman, D., Antani, S., Simpson, M.S., Thoma, G.R.: Design and development of a multimodal biomedical information retrieval system. Journal of Computing Science and Engineering 6(2), 168–177 (2012)
9. Grompone von Gioi, R., Jakubowicz, J., Morel, J.M., Randall, G.: LSD: a Line Segment Detector. Image Processing On Line 2, 35–55 (2012)
10. García Seco de Herrera, A., Müller, H., Bromuri, S.: 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)
11. de Herrera, A.G.S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., Müller, H.: Overview of the imageclef 2013 medical tasks. In: Forner, P., Navigli, R., Tufis, D., Ferro, N. (eds.) Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013. CEUR Workshop Proceedings, vol. 1179. CEUR-WS.org (2013)
12. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., de Herrera, A.G.S., Tsikrika, T.: Overview of the CLEF 2011 medical image classification and retrieval tasks. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands. CEUR Workshop Proceedings, vol. 1177 (2011)
13. Licklider, J.C.R.: A picture is worth a thousand words: And it costs... In: Proceedings of the Joint Computer Conference. pp. 617–621. AFIPS '69 (Spring), ACM, New York, NY, USA (1969)
14. Lopez, L.D., Yu, J., Arighi, C.N., Tudor, C.O., Torii, M., Huang, H., Vijay-Shanker, K., Wu, C.H.: A framework for biomedical figure segmentation towards image-based document retrieval. BMC Systems Biology 7(S-4), S8 (2013)
15. Müller, H.: Medical (visual) information retrieval. In: Information retrieval meets information visualization, winter school book. Springer LNCS, vol. 7757, pp. 155–166 (2013)
16. Müller, H., de Herrera, A.G.S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., Eggel, I.: Overview of the imageclef 2012 medical image retrieval and classification tasks. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012. CEUR Workshop Proceedings, vol. 1178 (2012)
17. Müller, H., Michoux, N., Bandon, D., Geissbühler, A.: A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. I. J. Medical Informatics 73(1), 1–23 (2004)
18. Murphy, R.F., Velliste, M., Yao, J., Porreca, G.: Searching online journals for fluorescence microscope images depicting protein subcellular location patterns. In: Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering. pp. 119–128. BIBE '01 (2001)
19. Rahman, M.M., You, D., Simpson, M.S., Antani, S., Demner-Fushman, D., Thoma, G.R.: Interactive cross and multimodal biomedical image retrieval based on automatic region-of-interest (ROI) identification and classification. Int. J. Multimed. Info. Retr. 3(3), 131–146 (2014)
20. Santosh, K.C., Antani, S., Thoma, G.: Stitched biomedical multipanel figure separation. In: International Symposium on Computer Based Medical Systems (2015)

21. Simpson, M.S., Demner-Fushman, D., Antani, S., Thoma, G.R.: Multimodal biomedical image indexing and retrieval using descriptive text and global feature mapping. Inf. Retr. 17(3), 229–264 (2014)
22. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., de Herrera, A.G.S., Bromuri, S., Amin, M.A., Mohammed, M.K., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., del Mar Roldán García, M.: General Overview of ImageCLEF at the CLEF 2015 Labs. Lecture Notes in Computer Science, Springer International Publishing (2015)
23. Yu, H.: Towards answering biological questions with experimental evidence: automatically identifying text that summarize image content in full-text articles. pp. 834–838 (2006)