AAUITEC at ImageCLEF 2015: Compound Figure Separation

Mario Taschwer[1] and Oge Marques[2]

[1] ITEC, Klagenfurt University (AAU), Austria, mario.taschwer@aau.at
[2] Florida Atlantic University (FAU), Boca Raton, FL, USA, omarques@fau.edu

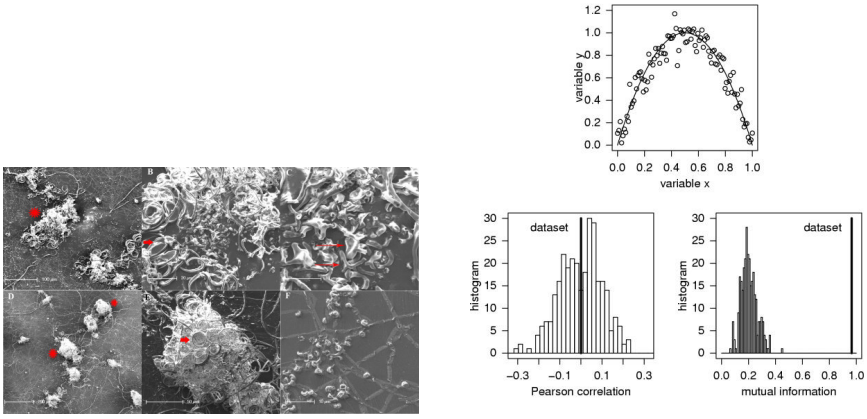
Abstract. Our approach to automatically separating compound figures appearing in biomedical articles is split into two image processing algorithms: one is based on detecting separator edges, and the other tries to identify background bands separating subfigures. Only one algorithm is applied to a given image, according to the prediction of a binary classifier trained to distinguish graphical illustrations from other images in biomedical articles. Our submission to the ImageCLEF 2015 compound figure separation task achieved an accuracy of 49% on the provided test set of about 3400 compound images. This stays clearly behind the best submission of other participants (85% accuracy), but is by an order of magnitude faster than other approaches reported in the literature.

1 Introduction

Automatically separating compound figures has been identified as a relevant problem for image-based information retrieval in collections of biomedical articles [1, 3, 5]. The task has been posed as a subproblem of the ImageCLEF 2015 [6] medical classification task [4]. Figure 1 shows two sample compound images of the provided training dataset.

Known approaches to the compound figure separation problem [1, 2] focus on the detection of homogeneous image regions separating subfigures, which we call *separator bands*. These approaches fail for compound images where subimages are stitched together without separator bands, as shown in Fig. 1(a). We therefore propose an approach based on edge detection that is able to separate subimages without separator bands, and is more generally applicable to subfigures whose rectangular border is represented by visible edges.

To handle subfigures not separated by vertical or horizontal edges, as shown in Fig. 1(b), we propose a variant of our algorithm which detects separator bands. The edge-based and band-based algorithms are applied selectively to a given compound image based on the prediction of a binary classifier trained to distinguish between graphical illustrations and other images. We assume that only graphical illustrations need to be handled by the band-based separation algorithm, whereas other compound images can be processed successfully by the edge-based algorithm.



(a) (b)

Fig. 1. Sample compound images (of the ImageCLEFmed 2013 dataset [3]) suitable for two different separator detection algorithms: (a) subimages are separated by vertical and horizontal edges, (b) subfigures are separated by whitespace (separator bands).

Although the proposed algorithm achieved only a moderate accuracy of 49% on the ImageCLEF 2015 test dataset, we believe that it may be useful for processing large image collections due to its efficiency. The average processing time per image (0.12 seconds) is about 20 times lower than the value reported by [2] for their approach. Moreover, separation performance is likely to be increased by incorporating additional techniques found to be effective like image markup removal or image label extraction [1].

The paper is organized as follows. Section 2 describes our approach in detail, results of the experimental evaluation at ImageCLEF 2015 are presented in Section 3, and Section 4 provides some ideas for future work.

2 Approach

Our approach to compound figure separation is a recursive algorithm (see Fig. 2) comprising the following steps: (1) classification of the compound image as illustration or non-illustration image, (2) removal of border bands, (3) detection of separator lines, (4) decision about vertical or horizontal separation, and (5) separation and recursive application to each subfigure image. The *illustration classifier* is used to decide which of two separator line detection modules to apply: if the compound image is classified as an illustration image, the *band-based* algorithm is applied, which aims at detecting separator bands between subfigures. Otherwise, the image is processed by the *edge-based* separator detection algorithm, which applies edge detection and Hough transform to locate candidate separator edges. The algorithm selection is based on the assumption that only compound images of graphical illustrations have no visible vertical or horizontal edges separating subfigures. The following four sections describe

the illustration classifier, the main recursive algorithm, and the two separator detection modules in more detail.

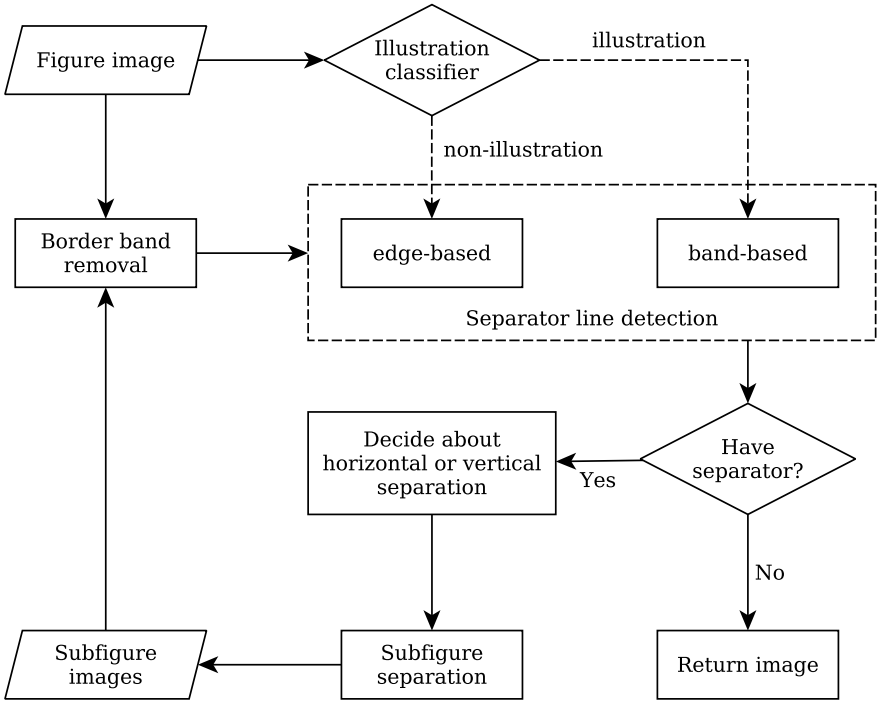


Fig. 2. Recursive algorithm for compound figure separation.

2.1 Illustration Classifier

In order to automatically discriminate between graphical illustrations and other images in the dataset, a logistic regression classifier has been trained on the training dataset of the ImageCLEF 2015 multi-label image classification task. The training dataset consists of about 1000 images of 29 classes (organized in a class hierarchy), which have been aggregated into two meta classes for the purpose of training the illustration classifier: the *illustration* meta class comprises all "general biomedical illustration" images of the training dataset except for chromatography images, screenshots, and non-clinical photos. Images of the latter classes and all diagnostic images have been assigned to the *non-illustration* meta class. About 36% of the images in the training set are labeled with multiple classes (compound images); for assignment to meta classes, we used only the first label and ignored all other labels.

We use just two simple global image features as classifier input, computed after gray-level conversion: *entropy*, estimated using a 256-bin histogram, and *mean intensity*. Classification performance has been evaluated on the test dataset of the ImageCLEF 2015 multi-label image classification task where ground truth annotations were assigned to illustration and non-illustration meta classes as described above. The test dataset contains about 500 images where about 44% are compound images. The accuracy of our illustration classifier on the test dataset was measured as 82.5% (92.0% on training dataset due to linear decision boundary).

The illustration classifier is used to decide which separator detection algorithm to apply to a given compound image. If the image is predicted to be an illustration with probability $p > 0.5$, the band-based separator detection is applied, otherwise the edge-based separator module is used. This decision is made only once for each compound image, so all recursive invocations use the same separator detection algorithm.

2.2 Recursive Algorithm

Before applying the main algorithm (Fig. 2) to a given compound figure image, it is converted to 8-bit gray-scale. *Border band removal* detects a rectangular bounding box surrounded by a maximal homogeneous image region adjacent to image borders (border band). The *separator line detection* modules return two lists of vertical and horizontal separator lines, respectively. If both lists are empty, recursion is terminated and the current image (without border bands) is returned. Due to minimal distances between separator lines and, additionally, to borders, recursion is guaranteed to terminate by finding no more separator lines at some point. The *decision about vertical or horizontal separation* is trivial if one of both lists of separator lines is empty. Otherwise the decision is made based on the regularity of separator distances: locations of separator lines and borders are normalized to the range [0,1], and the direction (vertical or horizontal) yielding the lower variance of adjacent distances is chosen. The final step is *subfigure separation and recursion*. The current figure image is divided into subimages along the chosen separation lines, and the algorithm is applied recursively to each subimage.

2.3 Edge-based Separator Detection

One of the two alternatives for separator line detection is the edge-based algorithm, which aims at detecting full-length vertical or horizontal edges in a given gray-scale image without border bands. The separator line detection module is invoked separately for vertical and horizontal directions, so the algorithm deals with a single edge direction only, which we denote by θ.

Figure 3 gives an overview of the edge-based separator detection algorithm. The core components are unidirectional edge detection (Sobel filter) and peak selection in the one-dimensional Hough transform of the binary edge map. The

Hough transform counts the number of edge points aligned on each line in direction θ. So the peaks correspond to the longest edges in this direction, and their locations identify candidate separator edges. Candidate edges are filtered by a similar regularity criterion as used for deciding about vertical or horizontal separation (see Section 2.2), and consolidated by filling small gaps between edge line segments. Finally, edges that are too short in comparison to image height or width, or too close to borders are discarded.

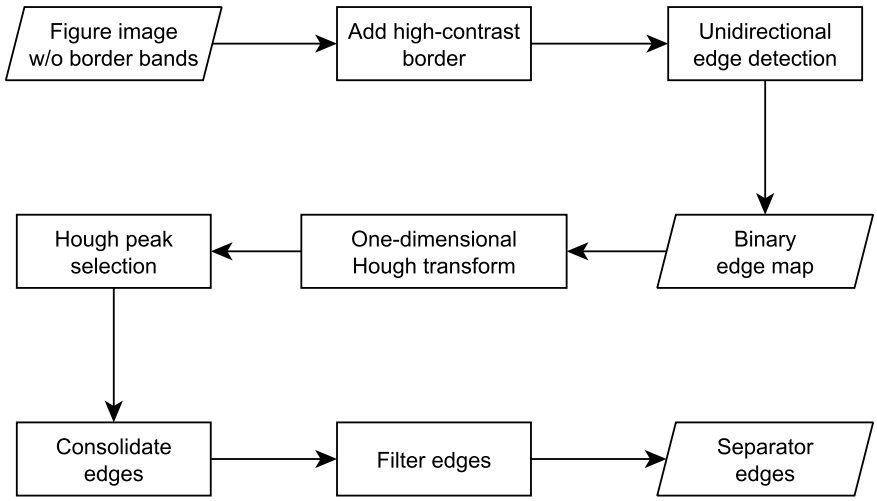


Fig. 3. Flow chart of edge-based separator line detection.

2.4 Band-based Separator Detection

For images without visible edges separating subfigures, an alternative separator detection approach is used. It aims at locating homogeneous rectangular areas covering the full width or height of the image, which we call *separator bands*. The steps of the proposed algorithm are depicted in Fig. 4.

Since band-based separator detection is intended to be applied to graphical illustration images, we binarize the image (using the mean intensity value as a threshold) and look for separator bands within white pixels only. We then compute mean projections along direction θ, that is, the mean value of each vertical or horizontal line of the binary image. A resulting mean value will be 1 (white) if and only if the corresponding line contains only white pixels. Candidate separator bands are then determined by identifying maximal runs of ones in the vector of mean values, and subsequently filtered using a regularity criterion similar to Hough peak selection (see Section 2.3). Finally, selected bands that

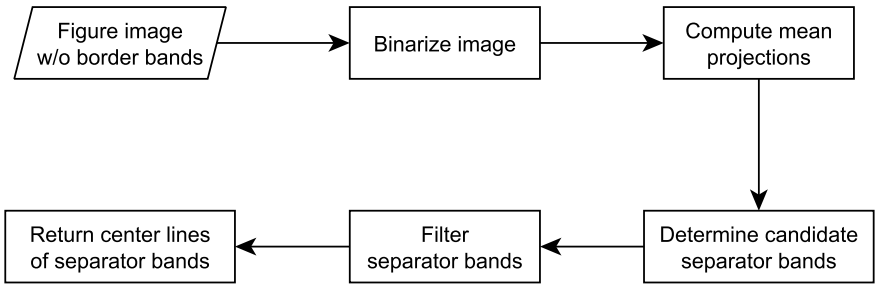


Fig. 4. Flow chart of band-based separator detection. The algorithm terminates early if no candidate separator bands are found (not shown).

are close to the image border are discarded, and the center lines of remaining bands are returned as separator lines.

3 Evaluation

Three runs have been submitted for the ImageCLEF 2015 compound figure separation task: (1) *aauitec_figsep_edge*: the proposed algorithm with edge-based separator detection was applied to all test images, (2) *aauitec_figsep_band*: only the band-based separator detection was used, and (3) *aauitec_figsep_combined*: edge-based or band-based separator detection was selected using the illustration classifier as described in Section 2.1. Runs (1) and (2) did not use the illustration classifier.

The test set contains 3381 compound images, of which 1839 (54%) have been classified as illustration images by our classifier. Our algorithm was implemented in Matlab and executed on a PC with 8 GB RAM and an Intel E8400 CPU running at 3 GHz. Experimental results are shown in Table 1.

Table 1. Experimental results on the ImageCLEF 2015 compound figure separation test set. Accuracy was measured using the official evaluation procedure. The best run was submitted by the NLM group (U.S. National Library of Medicine).

Run	**Accuracy**	**Run time per image**
aauitec_figsep_band	30%	46 ms
aauitec_figsep_edge	35%	157 ms
aauitec_figsep_combined	49%	117 ms
best submission	85%	

The combined approach using the illustration classifier shows a substantial improvement in detection accuracy compared to the other two variants of our

algorithm. This fact indirectly verifies our assumption that band-based separator detection is better suited for graphical illustrations than for non-illustration images.

The run time reported in Table 1 is the average run time per image when executed once for all (about 3400) images in the test set. The processing rate of about 9 images per second for the combined algorithm indicates that the algorithm may be applicable to large image collections if optimized and ported to C++. Note that the efficiency of other known approaches in the literature is either not documented [1] or by an order of magnitude lower ([2] reported 2.4 seconds per image).

Figure 5 shows some examples of test images where our algorithm failed for different reasons. Low-contrast edges present a problem for the edge-based algorithm, because they may appear too short compared to image height (or width). Errors of the illustration classifier may lead to the application of an inappropriate separator line detection method. The band-based algorithm fails if separator bands are cluttered with text that prevents detection of full-length white bands. Under-segmentation may occur for the band-based algorithm if separator bands are too thin; this problem may be alleviated by parameter optimization in further work. Isolated image labels may be detected as separate subfigures, both by edge-based and band-based algorithms, leading to over-segmentation. Note that the effect of border band removal in Fig. 5(e) does not reduce the score computed by the ImageCLEF evaluation procedure.

4 Conclusion and Further Work

We presented a recursive image processing algorithm for automatic separation of compound figures appearing in scientific articles. The algorithm has been evaluated on a dataset of compound images taken from biomedical articles in the context of the ImageCLEF 2015 compound figure separation task. Although the achieved detection accuracy of 49% is clearly inferior to the best result submitted by competitors (85%), early qualitative evaluation suggests that our algorithm provides benefits in terms of run-time efficiency and spatial detection accuracy. A quantitative evaluation will be the subject of future work.

Moreover, there is some potential for improving detection performance by modifying and extending the proposed algorithm. Firstly, internal parameters of the algorithm can be optimized using a cross-validation dataset. In our current implementation, parameters were set manually, as the evaluation tool was not available during development. Secondly, the quality of the training set for the illustration classifier can be improved by using all image labels of the multi-label image classification training set. Thirdly, the illustration classifier can be applied to every detected subfigure in order to select the separator detection algorithm on each recursive invocation (not just once for the entire compound image). Finally, additional image processing steps known to help figure separation could be added. Promising candidates include image markup removal and image label extraction [1].

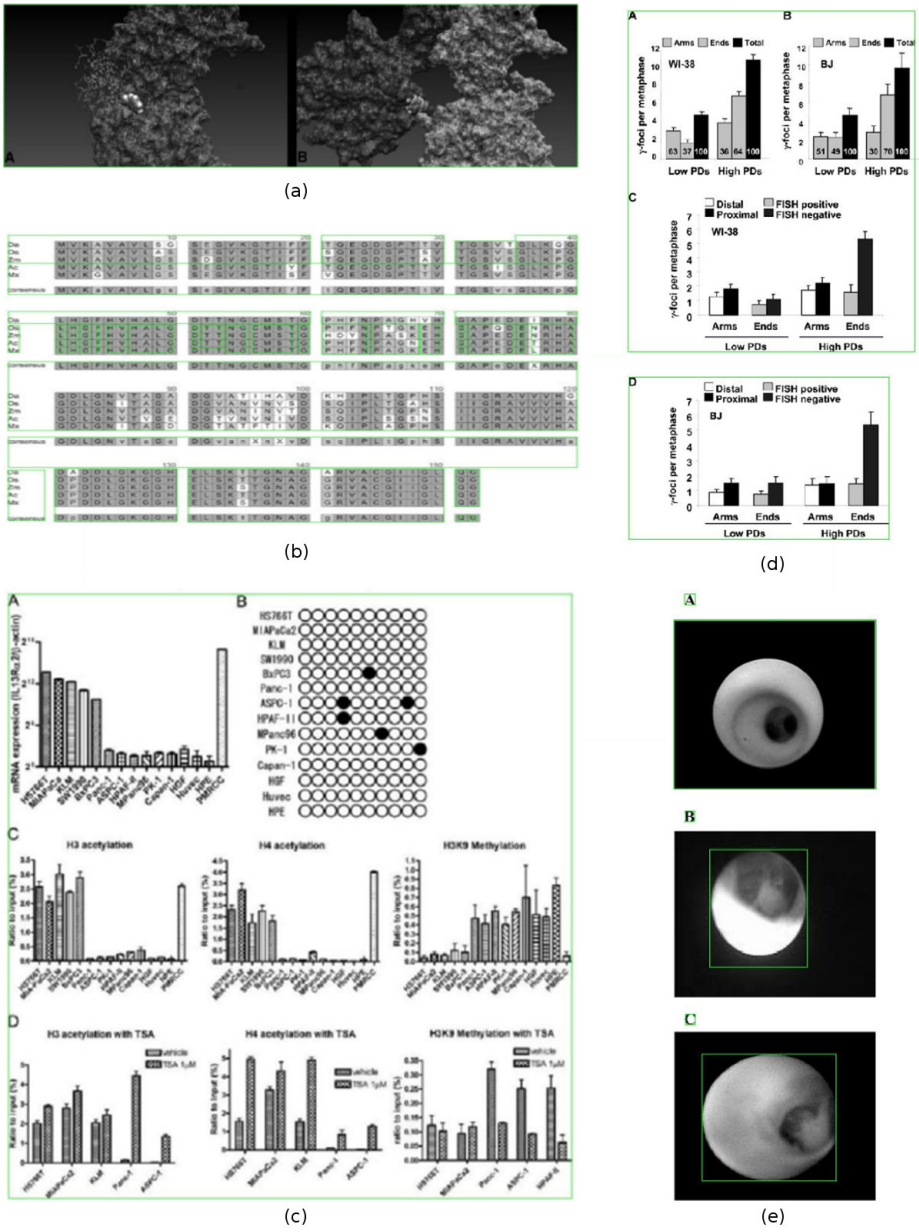


Fig. 5. Sample images of the compound figure separation test dataset [4] where our algorithm failed: (a) edge-based algorithm failed due to low-contrast edges; (b) illustration image processed by edge-based algorithm due to classification error (false negative of illustration classifier); (c) band-based algorithm failed due to text clutter; (d) under-segmentation by band-based algorithm due to thin separator bands; (e) over-segmentation (here by edge-based algorithm) due to isolated image labels.

References

1. 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), http://dx.doi.org/10.1002/asi.22810
2. Chhatkuli, A., Foncubierta-Rodríguez, A., Markonis, D., Meriaudeau, F., Müller, H.: Separating compound figures in journal articles to allow for subfigure classification. Proc. SPIE 8674, 86740J–86740J–12 (2013), http://dx.doi.org/10.1117/12.2007897
3. García Seco de Herrera, A., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., Müller, H.: Overview of the ImageCLEF 2013 medical tasks. In: CLEF 2013 Working Notes. CEUR Workshop Proceedings, ISSN 1613-0073, vol. 1179 (September 2013), http://ceur-ws.org/Vol-1179/
4. García Seco de Herrera, A., Müller, H., Bromuri, S.: Overview of the ImageCLEF 2015 medical classification task. In: CLEF 2015 Working Notes. CEUR Workshop Proceedings, ISSN 1613-0073, vol. 1391 (September 2015), http://ceur-ws.org/Vol-1391/
5. Simpson, M.S., You, D., Rahman, M.M., Xue, Z., Demner-Fushman, D., Antani, S., Thoma, G.: Literature-based biomedical image classification and retrieval. Computerized Medical Imaging and Graphics 39, 3–13 (2015), http://www.sciencedirect.com/science/article/pii/S0895611114000998
6. 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. In: Sixth International Conference of the CLEF Association, CLEF'15, Toulouse, September 8-11, 2015, Lecture Notes in Computer Science, vol. 9283. Springer International Publishing (2015)