Integration of Text Information and Graphic Composite for PDF Document Analysis

Canhui Xu[1,2,3], Zhi Tang[1,3], Xin Tao[1,3], and Cao Shi[1]

[1] Institute of Computer Science and Technology, Peking University, Beijing, China
[2] Postdoctoral Workstation of the Zhongguancun Haidian Science Park and Peking University Founder Group Co. Ltd, Beijing, China
[3] State Key Laboratory of Digital Publishing Technology, Beijing, China
{ccxu09,caoshi}@yeah.net, tommie@founder.com.cn, jolly.tao@pku.edu.cn

Abstract. The trend of large scale digitization has greatly motivated the research on the processing of the PDF documents with little structure information. Challenging problems like graphic segmentation integrating with texts remain unsolved for successful practical application of PDF layout analysis. To cope with PDF documents, a hybrid method incorporating text information and graphic composite is proposed to segment the pages that are difficult to handle by traditional methods. Specifically, the text information is derived accurately from born-digital documents embedded with low-level structure elements in explicit form. Then page text elements are clustered by applying graph based method according to proximity and feature similarity. Meanwhile, the graphic components are extracted by means of texture and morphological analysis. By integrating the clustered text elements with image based graphic components, the graphics are segmented for layout analysis. The experimental results on pages of PDF books have shown satisfactory performance.

Keywords: PDF document, graphic segmentation, graph based method, text clustering.

1 Introduction

Large-scale digitization projects undergoing at public and commercial digital libraries, such as the Million Book Project and the Google Book Search database, have indicated the significance and necessity of large scale processing of electronic documents. Different from scanned documents, the born-digital documents are generated by document processing software such as Microsoft Word, PowerPoint and LaTex. In addition, it has reliable typesetting information, such as embedded style and font information for textual content. However, current digitalization and OCRed format like Portable Document Format (PDF) documents contain no logical structure at any high level, such as explicitly delimited paragraphs, captions, or figures. By using formatting and font features embedded within the document, identifying logical structure of the document has attracted much attention both in academic and practical fields.

M. Zhou et al. (Eds.): NLPCC 2012, CCIS 333, pp. 13–22, 2012.

As the premise of robustness of logical layout understanding, various researches on layout analysis of PDF format documents were launched [1-2]. ICDAR (International Conference on Document Analysis and Recognition) has already held two competitions on Book Structure Extraction Competition focusing on structure recognition and extraction for digitized books [3-4]. It is well known that text only documents like novels are relatively easy to handle for PDF converters [1]. In applications of converting PDF to re-flowable formats like ePub or CEBX (Common e-Document of Blending XML), reliable layout analysis is highly desired to enrich the reading experience of e-book on small portable screens of handheld devices, such as mobile phone and PDA. Due to the large variety of the documents categories, open problems remain challenging for reliable layout analysis of PDF converters, including graph recognition integrating with text segmentation [1], tables and equation identification, etc..

Current page segmentation methods participating in ICDAR Page Segmentation Competitions perform better in separating text than non-text regions [5]. However, it is claimed that the leading text segmentation algorithms still have limitations for contemporary consumer magazine [2]. Illustrative graphic segmentation receives little attention. A complete layout understanding system requires that the full reconstruction of the document in scale of both high semantic and low-level [6]. For this purpose, the graphics in documents need to be segmented and identified accurately.

2 Related Work

Image-based document analysis and understanding has been discussed for decades. Most of the existing research concentrates on inputs objects like scanned images or camera documents. Image-based document layout analysis segments the document image into homogenous geometric regions by using features like proximity, texture or whitespace. Most of the research has been done on connect components (CCs) of page images. The “docstrum” method [7] exploited the k nearest-neighbor pairs between connect component centers by features like distance and angle. Kise [8] pointed out that connected components of black pixels can be utilized as primitives so as to simplify the task of page segmentation by combining connected components appropriately. It performed page segmentation based on area Voronoi diagrams by using distance and area ratio of connected components for deleting superfluous edges. Simon [9] proposed a bottom-up method based on Kruskal algorithm to classify the text and graph. Xiao [10] utilized a Delaunay triangulation on the point set from the bounded connected components, and described the page structure by dividing the Delaunay triangles into text area and fragment regions. Ferilli [11] used the distance between connect component borders for bottom-up grouping method. Recently, Koo [12] developed a new approach to assign state estimation on CCs to perform text block identification and text line extraction. It claims that the limitation of this method suffers from non-text objects.

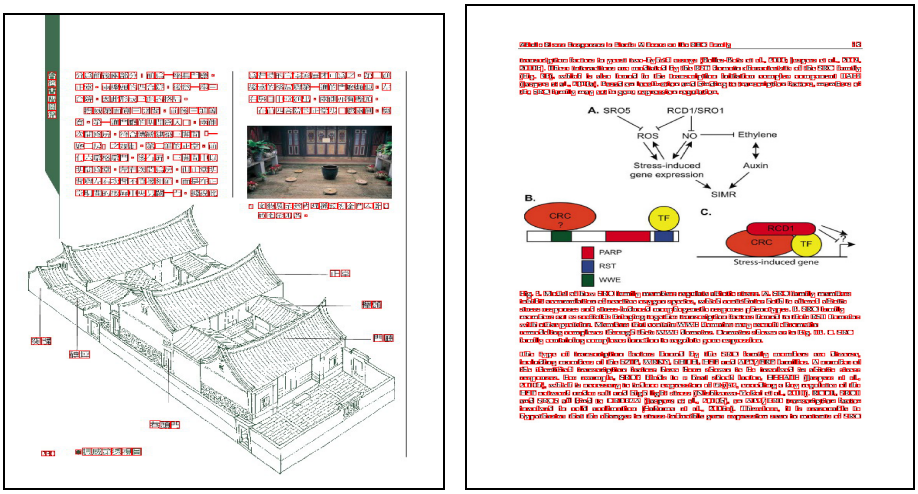
The methodologies in page segmentation can be extended for digital-born documents analysis such as PDF documents. A large number of documents are created or converted in PDF format. These documents represent characters and images in

explicit form, which can be straightforwardly exploited for layout analysis. Graphic, also called as figure or illustration in certain context, is a powerful way of illustrating and presenting the key ideas or findings. It has various categories such as photograph from conventional cameras or microscopes, drawing, 2D or 3D plot, diagram and flow chart. In PDF, figures and tables usually need to be recognized through grouping page primitives such as lines, curves, images and even text elements. Current digital library metadata has little improvement in graphic identification covering all kinds of documents. There exist several attempts in graphic identification and understanding. Chao [13] proposed a method to identify and extract graphic illustrations for PDF documents, which is based on the proximity of page elements. Shao [14] focused the research on graphic recognition of figures in vector-based PDF documents by using machine learning to classify figures with various grapheme statistics, which aims at the application of diagram retrieval. However, the extraction of the figure content including graphics and the text inside the figures is accomplished by grouping the primitives near the end of the content stream in PDF articles published with a standard Adobe FrameMaker template, which is not the common case in most of PDF books or magazines. The Xed system [15] is proposed to convert PDF to XML, and it claimed that traditional document analysis will drastically improve PDF's content extraction.

It is reported that the detection of graphic components and their integration with text segmentation will greatly improve the layout analysis performance [1]. In this paper, the goal is to group text into visually homogeneous regions and recognize the graph object integrating with text elements. A hybrid method is proposed and its application on PDF documents is presented. The preprocessing step and the graph based method are presented in Section 3 and 4. Its application on PDF sources is presented at section 5. The conclusion is given in Section 6.

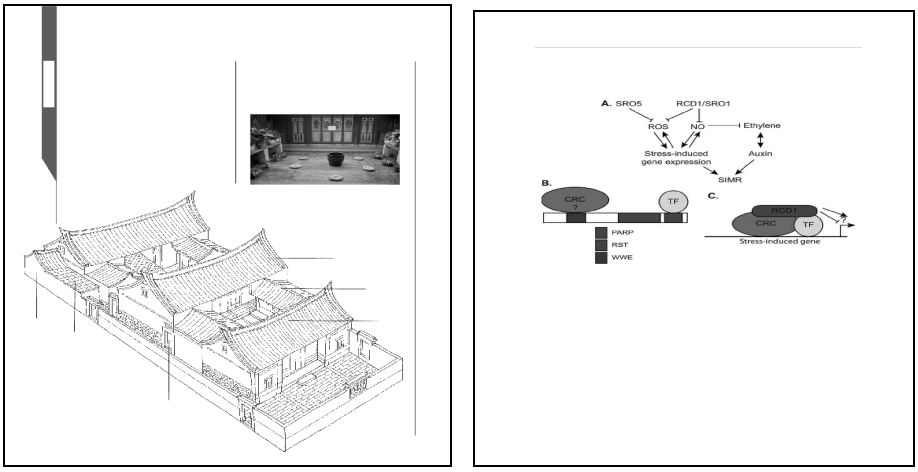
3 Preprocessing

It is assumed that the inherent meta-data structure information can be provided by born-digital documents like true PDFs, other than PDF files embedded with scanned document page image. The PDF documents are described by low-level structural objects like text elements, lines, curves and images etc., and associated style attributes such as font, color, etc.. All the basic low-level elements from a PDF document are extracted in preprocessing step. The PDF parser using in this work is provided by Founder Corporation to parser the low-level objects, which is introduced in [16]. Basic objects here imply the elements or primitives in each page, which cannot be subdivided into smaller objects, including text, image elements or operations. The composite object is constituted by basic objects with certain predefined similarity. A graphic object includes picture, geometric graphic or graphic character element. From another perspective, text elements can be categorize as body text from the article, text belonging to graph or picture (or image), text inside table, and text for footer, header or other decorations, etc..



(a) (b)

Fig. 1. PDF document pages with bounding boxes on the parsed text elements



(a) (b)

Fig. 2. Non-text objects in the “left-over” PDF document images

The complexity of graphic recognition has become a significant step to be handled in building a complete document understanding system. To extract the graphic and text primitives corresponding to figures, a hybrid method incorporating both low-level structural elements analysis and vision based image analysis is proposed. Firstly, the bounding boxes of the parsed text elements embedded by PDF are exported and then converted from the metric of logic units to pixels. As Fig.1 has demonstrated, by imposing the bounding boxes of text elements on the original document page image, the

super-pixel representation of a page image has provided the layout analysis with great convenience. The text regions in a page image can be regarded as an array of bounding boxes in two dimensions. As is given in Fig.1 (a), a two-column page in Chinese from an electronic book is plotted with all the bounding boxes of the parsed text elements. It contains a composite graph combining a line graph component with surrounded text elements, a photo graph, horizontal body text and vertical marginal decoration, etc. Fig.1 (b) is an example page from an English e-book crawled from the web. These cases are challenging to segment the graphic components when either the graphic has been embedded with text or has touched text elements. For PDF documents, in fact, the graphic of line drawing in Fig.1 (a) is not parsed as a whole object but as numerous paths and sub-images, which is the same case for the flowchart graphic in Fig.1 (b).

To use the layer information provided by low-level structure elements embedded in PDF documents, the extracted text elements in each page are subtracted from the original input image before passing the cleaned image to graphic region analysis. As is shown in Fig.2, all the text elements in pages are covered with white pixels and the non-text page images only contains graphic parts, lines and decorations, etc..

4 Graphic Segmentation

After the preprocessing step, the text layer input is analyzed by applying the graph based analysis proposed in section 4.1, and the non-text graphic layer is processed by texture features and morphological analysis given in section 4.2.

4.1 Graph Based Analysis

As perceptual grouping works in human vision, graph-based method [17] developed can capture certain perceptually important non-local image characteristics for segmentation purposes. In [12], the connect component (CCs) are used as graph vertices. Unlike its application on image segmentation in pixel level or CC state, in this paper, page element or primitive corresponding to a vertex are constructed in the graph. All the text elements can be connected by establishing a neighbourhood system. Delaunay tessellation is applied in this regard. It is a convenient and powerful neighbourhood representation of 2D image.

An undirected graph can be defined as $G=(V,E)$ whose vertex set is V and $(v_i,v_j)\in E$ are the edges connecting two vertexes. The dissimilarity between adjacent elements v_i and v_j is measured as weights $w(v_i,v_j)$ for each edge $(v_i,v_j)\in E$ constructed. In this application, the elements in V are the centroids of the bounding boxes extracted from PDF parser.

$$w(v_i,v_j)=\sum_k \lambda_k f_k(v_i,v_j) \tag{1}$$

where k is the dimension of feature dissimilarity $f_k(v_i, v_j)$ between adjacent elements v_i and v_j, and λ_k is the coefficient corresponding to each feature function. $w(v_i, v_j)$ is a linear combination of the selected feature functions. The undirected graph constructed can be called as page graph in the field of document analysis [9].

Two feature functions are defined based on the Euclidean distance function $f_E(v_i, v_j)$ and an angle dissimilarity function $f_A(v_i, v_j)$:

$$f_E(v_i, v_j) = \left[(v_i(x) - v_j(x))^2 + (v_i(y) - v_j(y))^2 \right]^{1/2} \tag{2}$$

$$f_A(v_i, v_j) = \left[\tan^{-1} \frac{\Delta y_{i,j}}{\Delta x_{i,j}} \right]_{180^\circ} \tag{3}$$

where $\Delta x_{i,j} = \left| v_j(x) - v_i(x) \right|$, $\Delta y_{i,j} = \left| v_j(y) - v_i(y) \right|$, $[\cdot]_{180^\circ}$ indicates $0 \le f_A(v_i, v_j) \le 180^\circ$. The weight for each edge $(v_i, v_j) \in E$ is selected as:

$$w(v_i, v_j) = \lambda_E f_E(v_i, v_j) + \lambda_A f_A(v_i, v_j) \tag{4}$$

As are defined in [17], the internal difference $Int(C)$ within one composite component $C \subseteq V$ and the inter-component difference $Dif(C_1, C_2)$ between two components $C_1, C_2 \subseteq V$ play an important role in graph based segmentation, which are formulated as:

$$Int(C) = \max_{e \in MST(C,E)} w(e) \tag{5}$$

$$Dif(C_1, C_2) = \min_{v_i \in C_1, v_j \in C_2, (v_i, v_j) \in E} w((v_i, v_j)) \tag{6}$$

The measures indicate that when the edge weights are equal or smaller than $Int(C)$, the vertices connecting them are considered to be in one composite component. If two vertices are not in the same component, $Dif(C_1, C_2)$ is larger than the internal difference within at least one components, $Int(C_1)$ and $Int(C_2)$. The pairwise region comparison predicate is defined as:

$$D(C_1, C_2) = \begin{cases} true, & if \quad Dif(C_1, C_2) > MInt(C_1, C_2) \\ false, & otherwise \end{cases} \tag{7}$$

where the minimum internal difference $MInt(C_1, C_2)$ is defined as:

$$MInt(C_1, C_2) = \min(Int(C_1) + \tau(C_1), Int(C_2) + \tau(C_2)) \tag{8}$$

To identify an evidence for a partition, the difference between two components must be greater than the internal difference. The extreme case is that the size of component is 1, and $Int(C) = 0$. Therefore, a threshold function τ can solve this problem:

$$\tau(C) = 1/|C| \tag{9}$$

where $|C|$ is the size of C.

In the graph based method, the edge causing the grouping of two components is exactly the minimum weight edge between the components. That implies the edges causing merges are exactly the edges that would be selected by Kruskal's algorithm for constructing minimum spanning tree of each component [17]. The computational complexity is reduced by path-compression.

A spanning tree of a page graph is defined as a tree contains all the vertices of a graph, which indicates that when given n_V vertices or primitives in a page, the spanning tree of the page has $n_V - 1$ edges. In the undirected graph $G = (V, E)$, the goal is to find an acyclic subset $F \subseteq E$ connecting all the vertices. And the total weight is minimized:

$$w(F) = \sum_{(v_i, v_j) \in F} w(v_i, v_j) \tag{10}$$

Minimal spanning tree requires that the sum of the edge weights is minimal among all other possible spanning trees of the same graph. A minima spanning tree of page graph is built by the Kruskal algorithm.

4.2 Texture Entropy and Morphological Analysis

The co-occurrence matrix is generally defined over an image. It reflects the distribution of co-occurring values at a given offset. Mathematically, it is formulated as:

$$C_{\Delta x, \Delta y(i,j)} = \sum_{p=1}^{n} \sum_{q=1}^{m} \begin{cases} 1, & \text{if } I(p,q) = i \text{ and } I(p + \Delta x, q + \Delta y) = j \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where C is a co-occurrence matrix over image I with size $n \times m$, and $(\Delta x, \Delta y)$ is the offset. The texture entropy En is defined as:

$$En = \sum \sum p_{ij} \log p_{ij} \tag{12}$$

where i and j are the row and column, p_{ij} is the probability matrix.

The morphological filter consisting of opening followed by closing is applied to eliminate the noise, and region filling is performed on the preprocessed page image containing non-textual graphic objects. The outside bounding box of graphic object can be identified on the specific connected component.

5 Graphic Segmentation Results and Analysis

5.1 Delaunay Triangulation and Text Elements Clustering

To construct the neighborhood system of text elements, Delaunay triangulation is generated by conventional incremental method. The features from all the vertices of triangles are extracted, including font dissimilarity, Euclidean distance, orientation angle, which can be utilized for further segmentation. In these two cases, the text elements are clustered into the right graphic regions.

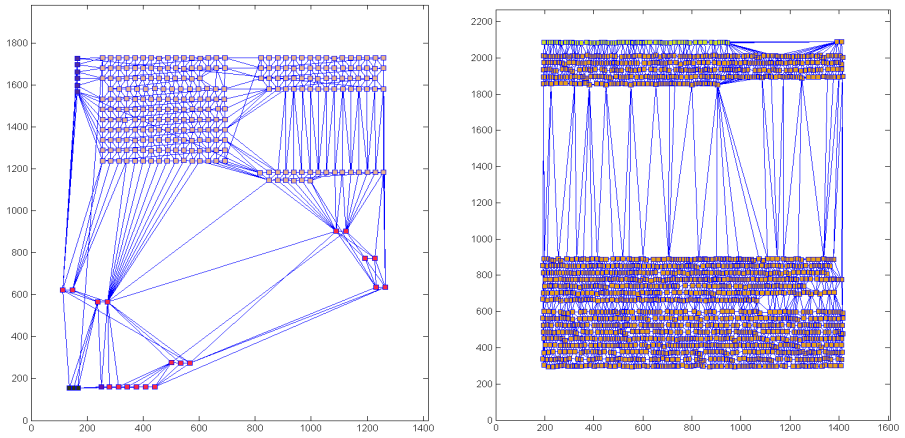


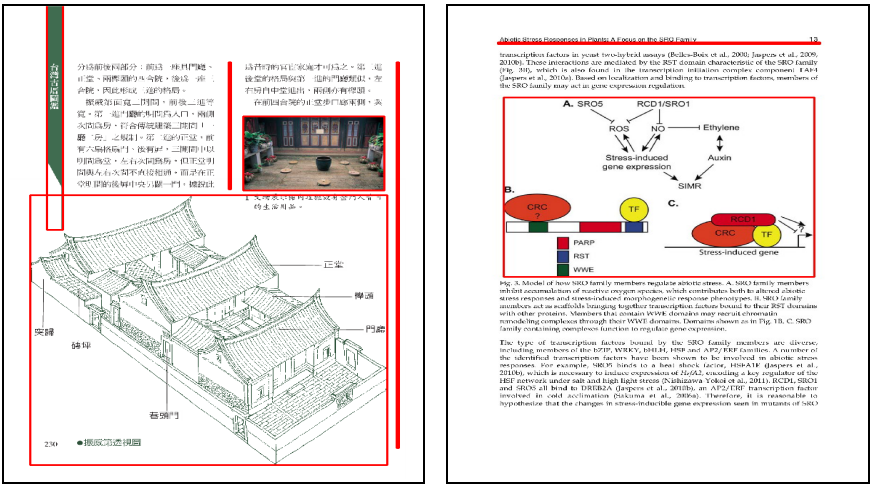
Fig. 3. Delaunay triangulation tessellation of page text elements and graph based partition

As can be seen, the super-pixel representation and Delaunay triangulation of Fig.1 (a) and (b) are illustrated in Fig.3 (a) and (b) respectively. The clustering of text elements is based on the algorithm proposed in Section 4.1 according to predefined threshold of feature dissimilarity, which is a combination of font similarity and Euclidean distance. The clusters of the elements are presented in different marker face color. The text elements belonging to the body text area, title, foots and notes, graphic components are clustered into separate classes. Similarly, the outside bounding boxes of each class can be identified on the page images for the purpose of integrating text elements and graphic object.

5.2 Segmentation of Graphics

The graphic segmentation results in Fig.4 are satisfactory. By region grouping technique, the line drawing in Fig.4 (a) of an architecture integrating with surrounded illustrative text elements is accurately detected, so are the photo graph, separating lines and marginal graph. The composite graph in Fig.4 (b) is identified as a whole component with all the pictorials texts.

The proposed hybrid segmentation algorithm was tested on two Chinese e-books, one English e-book and one consumer magazine. As is pointed out in [1], precise quantitative evaluation for books and magazines requires ground truth. Although the construction of evaluation set is very time-consuming, preliminary work has been already initiated, which will be further carried out in evaluation of both low-level and high-level page segmentation. However, we manually counted the integration of text and graph. It can achieve over 80% accuracy by means of counting the number of graphics.



(a) (b)

Fig. 4. Graphic segmentation results of PDF document pages in Fig.1

6 Conclusions

In this work, a hybrid segmentation scheme is proposed to segment the graphics in PDF documents. By utilizing inherent advantages of born-digital documents embedded with characters and images in explicit form, the provided structural information can benefit the layout analysis. Delaunay tessellation is applied on the centroids of the page elements to build the neighborhood system for parsed text elements. The proposed hybrid method uses graph based concept to group the text elements according to edge weights like the proximity and font information. Graphic segmentation integrating with texts is accomplished by text clustering and connected components segmentation. The experimental results on document pages of PDF books and magazines have shown satisfactory performance.

Acknowledgements. This work was supported by the National Basic Research Program of China (973 Program) (No. 2010CB735908).

References

1. Marinai, S., Marino, E., Soda, G.: Conversion of PDF Books in ePub Format. In: 11th International Conference on Document Analysis and Recognition, pp. 478–482 (2011)
2. Fan, J.: Text Segmentation of Consumer Magazines in PDF Format. In: 11th International Conference on Document Analysis and Recognition, pp. 794–798 (2011)
3. Doucet, A., Kazai, G., Dresevic, B., Uzelac, A., Radakovic, B., Todic, N.: Book Structure Extraction Competition. In: 10th International Conference on Document Analysis and Recognition, pp. 1408–1412 (2009)
4. Doucet, A., Kazai, G., Meunier, J.-L.: Book Structure Extraction Competition. In: 11th International Conference on Document Analysis and Recognition, pp. 1501–1505 (2011)
5. Antonacopoulos, A., Pletschacher, S., Bridson, D., Papadopoulos, C.: Page Segmentation Competition. In: 10th International Conference on Document Analysis and Recognition, pp. 1370–1374 (2009)
6. Tombre, K.: Graphics Recognition: The Last Ten Years and the Next Ten Years. In: Liu, W., Lladós, J. (eds.) GREC 2005. LNCS, vol. 3926, pp. 422–426. Springer, Heidelberg (2006)
7. O'Gorman, L.: The Document Spectrum for Page Layout Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(11), 1162–1173 (1993)
8. Kise, K., Sato, A., Iwata, M.: Segmentation of Page Images Using the Area Voronoi Diagram. Computer Vision and Image Understanding 70, 370–382 (1998)
9. Simon, A., Pret, J.C., Johnson, A.P.: A Fast Algorithm for Bottom-up Document Layout Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(3), 273–277 (1997)
10. Xiao, Y., Yan, H.: Text Region Extraction in a Document Image Based on the Delaunay Tessellation. Pattern Recognition 36, 799–809 (2003)
11. Ferilli, S., Biba, M., Esposito, F.: A Distance-Based Technique for Non-Manhattan Layout Analysis. In: 10th International Conference on Document Analysis and Recognition, pp. 231–235 (2009)
12. Koo, H.I., Cho, N.I.: State Estimation in a Document Image and Its Application in Text Block Identification and Text Line Extraction. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 421–434. Springer, Heidelberg (2010)
13. Chao, H.: Graphics Extraction in PDF Document. In: Document Recognition and Retrieval X, Santa Clara, CA, USA, vol. 5010, pp. 317–325 (2003)
14. Shao, M., Futrelle, R.P.: Recognition and Classification of Figures in PDF Documents. In: Liu, W., Lladós, J. (eds.) GREC 2005. LNCS, vol. 3926, pp. 231–242. Springer, Heidelberg (2006)
15. Hadjar, K., Rigamonti, M., Lalanne, D., Ingold, R.: Xed: A New Tool for Extracting Hidden Structures from Electronic Documents. In: International Workshop on Document Image Analysis for Libraries, pp. 212–224 (2004)
16. Fang, J., Tang, Z., Gao, L.: Reflowing-Driven Paragraph Recognition for Electronic Books in PDF. In: SPIE-IS&T International Conference of Document Recognition and Retrieval XVIII, vol. 7874, pp. 78740U-1–78740U-9 (2011)
17. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient Graph-Based Image Segmentation. International Journal of Computer Vision 59(2), 167–181 (2004)