Graphic Composite Segmentation for PDF Documents with Complex Layouts

Canhui Xu*[abc], Zhi Tang[abc], Xin Tao[a], Cao Shi[a]
[a]Institute of Computer Science & Technology of Peking University, Beijing, China, 100080; [b]State Key Laboratory of Digital Publishing Technology (Peking University Founder Group), Beijing, China, 100080; [c]Postdoctoral Workstation of the Zhongguancun Haidian Science Park, Beijing, China, 100871

ABSTRACT

Converting the PDF books to re-flowable format has recently attracted various interests in the area of e-book reading. Robust graphic segmentation is highly desired for increasing the practicability of PDF converters. To cope with various layouts, a multi-layer concept is introduced to segment graphic composites including photographic images, drawings with text insets or surrounded with text elements. Both image based analysis and inherent digital born document advantages are exploited in this multi-layer based layout analysis method. By combining low-level page elements clustering applied on PDF documents and connected component analysis on synthetically generated PNG image document, graphic composites can be segmented for PDF documents with complex layouts. The experimental results on graphic composite segmentation of PDF document pages have shown satisfactory performance.

Keywords: PDF converter, graphic segmentation, complex layouts, image based document analysis

1. INTRODUCTION

Successful conversion of digital born documents such as Portable Document Format (PDF) to reflowable format like ePub highly depends on the premise of reliable layout analysis. Being different from image based documents layout analysis, the digitally originated documents like PDF documents are generated by document processing software such as Microsoft Word, PowerPoint or LaTex, and therefore meta-data structure information can be provided additionally. These low-level structural objects include text elements, lines, curves and images etc., and associated style attributes such as font, color, etc.. However, the generated PDF have only minimal low-level structure information. Moreover, no logical structure at any high level, such as explicitly delimited paragraphs, captions, or figures are contained. Recently, an increasing interest on the extraction of the logical structure of PDF files is reported both in academic and practical fields[1,2]. In applications like information retrieval[3] and reflowable document reconstruction[4], the provision of such structure information, for example XML format, can not only aid user navigation inside book and improve search performance but also profit the e-booking reading on handheld devices (e.g. smartphones, Kindle, iPad). Naturally, the research on PDF or other digital born documents is motivated by the urgent need of a system on e-book reading, which enables the automatic detection of document physical and logical structure. Pioneers of this field have made attempts on layout analysis [3], header and footer extraction [5,6], page body delimitation [7], paragraph recognition [4], table extraction[8] and mathematical formula identification[9]. A complete layout understanding system requires that the full reconstruction of the document in scale of both high semantic and low physical level[10]. Open problems remain challenging for reliable and effective PDF converters, including graph recognition integrating with text segmentation[11], Table of Contents (ToC), tables and equation identification, etc. And the robustness of layout analysis still has a large scope for improvement, especially in the cases of complex layout documents. For this purpose, the graphics in the documents need to be segmented and identified accurately.

The segmentation of text and graphics has been researched and reported during recent decades. However, image document layout analysis in many cases considers only the text objects segmentation, regarding the graphics as left-overs. Current page segmentation algorithms perform mostly better in the task of separating text than in segmenting non-text regions[12]. Most of the applications even consider the non-text regions exclusive to segmentation task for their application contexts. Still, there are some researches exploring text segmentation from mixed text/graphics[13,14] or

*ccxu09@yeah.net; phone 86-10-82179508;

Document Recognition and Retrieval XX, edited by Richard Zanibbi, Bertrand Coüasnon,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 8658, 86580E · © 2013 SPIE-IS&T
CCC code: 0277-786/13/$18 · doi: 10.1117/12.2003705

engineering drawings[15]. In Ref. 7, it claimed that there were roughly three approaches for separating graphic from text, including directional morphological filtering, extraction of lines and arcs, and connected component analysis. Like its popularity in the leading segmentation algorithms in ICDAR segmentation competitions, connected component analysis is most commonly selected to handle documents with complex layouts, but its parameters should be finely tuned based on the distinguish of spatial features between text and graphic objects.

There exist also several attempts in graphic segmentation in PDF documents. Usually, figures need to be extracted through grouping page primitives such as lines, curves, images and even text elements. Chao[16] proposed a method to identify graphic illustrations, which is based on the proximity of page elements. Shao[17] 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. For complex document layouts like contemporary consumer magazine, Fan[11] pointed out the graphic recognition and integration with text segmentation will greatly improve performance. Hadjar[3] pointed out that running traditional document analysis algorithms on PDF documents will drastically improve PDF's content extraction.

Well researched image-based documents layout analysis can be extended for PDF documents structure analysis. In this paper, instead of exploiting the visual image features of text objects, low level structure information of page text elements is provided by PDF parser. Hence, labeling pictorial connected components as text objects in image based layout analysis is of no problem for real digitally originated documents based layout analysis. A multi-layer based analysis is introduced to segment graphic composites by integrating the image based document analysis and inherent digital born document advantages. The background on image based and digital born based document analysis is introduced in Section 2. The preprocessing step and the multilayer based method are proposed in Section 3 and 4. Its application on PDF resources is presented at section 5. The conclusion is given in Section 6.

2. BACKGROUND

The objective of document structure analysis is to extract both physical and logical structure information, ranging from lowest level component to semantically highest level partition of document elements. There exist two correlated schemes for structure analysis: image based and digital born based for document analysis and understanding. The comparison and their procedure can be found in the block diagrams shown in Figure 1.

Traditional image based document structure analysis deals with inputs like image documents. The structure extraction is subjected to steps depicted in Figure 1 (a). The pre-processing, including noise removal and de-skewing, is to enhance the image quality for facilitating subsequent layout analysis. The existing layout analysis algorithms over three decades are generally specialized for certain applications. The variability of layouts makes it extremely difficult to have a general method performing good over all documents. After layout analysis, the extracted physical structure is preceded to OCR for character recognition. Finally, layout understanding is performed based on the outcome of physical structure and OCR-ed texts to derive logical structure.

In contrast to image based document structure analysis, Figure 1 (b) has inputs of digitized documents like PDF or OCR-ed secondary (ASCII) document representation texts. The advantages of the second scheme are that digital born documents can provide low level structure information like text, a partial amount of images, path operators and operations, and thus OCR or OFR is unnecessary during structure analysis procedure

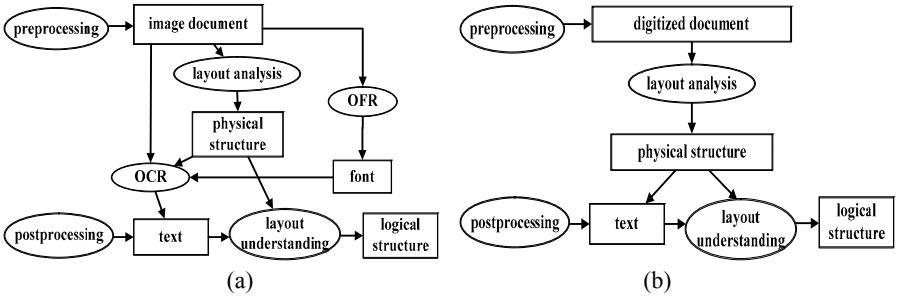


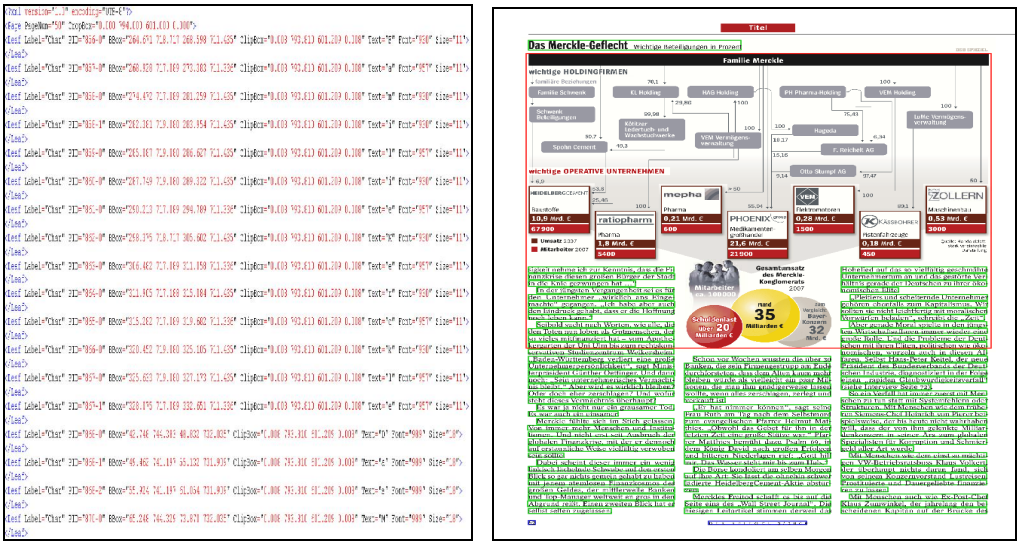
Figure 1. Comparison of two document structure analysis schemes: (a) traditional image based document structure analysis procedure; (b) digitized document structure analysis procedure.

To deal with the PDF documents with complex layouts, an effective and efficient approach must take all possible resource information, including: i) low-level page elements like text, image and path, ii) the reliable typesetting information provided by PDF parser, such as embedded style and font information for textual element, iii) the visual appearance of all page elements represented by rendered synthetic images. Hence, regarding each document page, there are three files for description:

1. A physical xml description of page elements and attributes. Three kinds of elements are discussed including text elements, image elements and path operations. Each element has its unique ID number, which can relate its physical label and logical label in further analysis of document structure reconstruction and reflowable format conversion. Figure 2 (a) gives an example of xml description of page Figure 2 (b).

2. A .png image with resolution of 300 dpi. This synthetic image can be rendered according to the selected page elements for specific application purpose.

3. A labeled ground-truth. It contains information for further performance evaluation, such as bounding boxes and element IDs. Figure 2 (b) is the visualization of manually labeled ground truth of text lines and graphic composite within one page. More details on this semi-automatic ground-truth tool can be found in our previous work[18].



(a) (b)

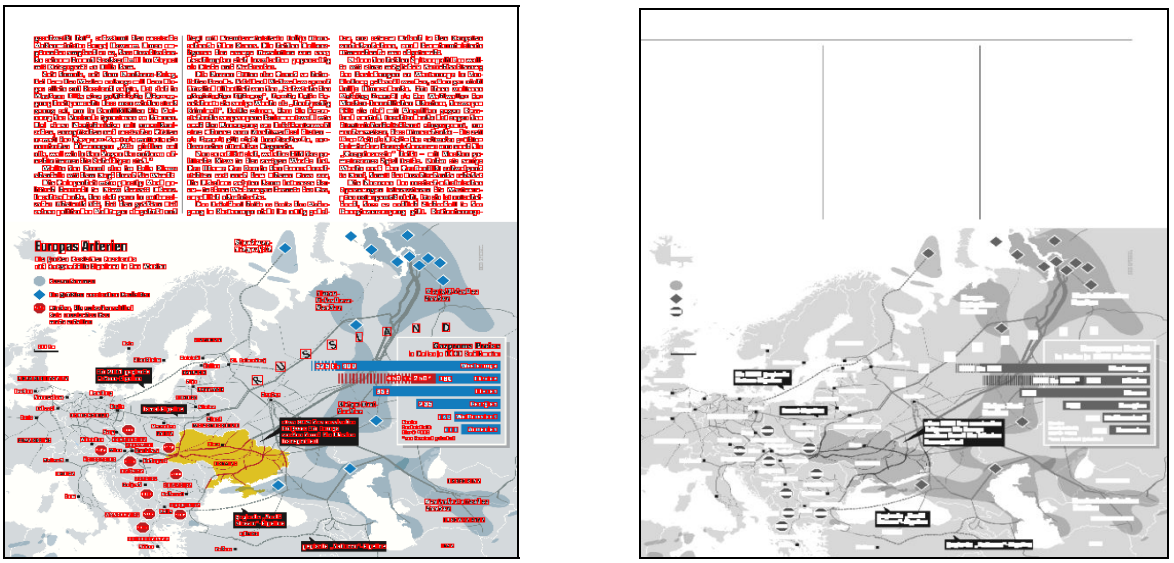
Figure 2. An Example of the descriptions for PDF document page: (a) a physical xml description of page elements and attributes; (b) a document page with labeled ground-truth.

3. PREPROCESSING

Graphic composites include picture or photographic images, geometric graphic or drawings. It is also called as illustrations in other application context. The extraction of the graphic composite covers graphic composites with text element insets as well as with nearby or surrounded text elements. From perspective of text elements, all extracted text elements within documents can be divided into body text, text belonging to graph or picture (or image), text inside table, and text for footer, header or other decorations.

With the assumption that the inherent meta-data structure information can be provided by true PDF, 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 indicate the elements or primitives in each page, which cannot be subdivided into smaller objects, including texts, images or path operations. The bounding boxes of page elements embedded by PDF are exported and then converted from the metric of logic units to pixels. Figure 3(a) is a three column page from the magazine “Der Spiegel, Mar. 2009”. As is demonstrated, by imposing the bounding boxes of text elements on the original document page image, this super-pixel representation of a page image has provided the layout analysis with great convenience. Each text primitive is presented as a rectangle bounding box with same size of a German letter. The text regions in a page image can be regarded as an array of

bounding boxes in two dimensions. In fact, the graphic of the whole map in Figure 3(a) is not parsed as a whole graphic object but as thousands of paths operations, which are the most cases among PDF document illustrations, which need great effects to process, especially the graphic composites embedded with or touched by text elements.



(a) (b)

Figure 3. (a) An example of PDF document pages covered with the text bounding boxes provided by PDF parser; (b) preprocessed image document with numerous graphic composites.

The extracted text elements are subtracted from the original input .png image before passing the preprocessed image to graphic region layout analysis. As is shown in Figure 3(b), all the text elements in pages are covered with white pixels and the non-text page objects only contains graphic parts, lines and other decorations, etc.

4. PROPOSED METHOD

To extract page primitives corresponding to a graphic composite, a multi-layer segmentation approach incorporating both structural representations and image based analysis is proposed. The page images are divided into text layer and non-text layer which are to be processed respectively, as is elaborated in the following subsections.

4.1 Analysis on Non-text based Layer

As is illustrated in Figure 3(b), the non-text based layer image involves only non-textual objects distributed over the page, with a premise that the text elements can be parsed with absolute accuracy. To segment this page image into regions corresponding to structural units, it is natural to consider the spatial arrangement of intensities or colors, which can be described by image texture features, since the graphic components (non-text objects) possess more local information content and details than the white colored background. In this non-text based layer analysis, two filters are applied, including local entropy to obtain the local measurement of information content and morphological filter to fill holes within each extracted connected graphic component.

Gray level co-occurrence matrices can capture properties of a texture. Given a gray-tone image I described by a set of intensity gray tones, a gray level co-occurrence matrix is a two-dimensional array C in which both the rows and the columns represent a set of possible image intensity values. The co-occurrence matrix is generally defined over an image. It reflects the distribution of co-occurring values at a given offset. The value $C(i, j)$ indicates the frequency of value i co-occurs with value j in pre-defined spatial relationship. Set Δ to be a displacement vector $(\Delta x, \Delta y)$, where Δx is a displacement in columns (horizontally to the right) and Δy is a displacement in rows (vertically downward).

Mathematically, it is formulated as:

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

where C is a co-occurrence matrix over image I with size $n\times m$.

To further analyze the properties of the texture, numeric features are computed from the co-occurrence matrix, which can be used to represent the texture more compactly. General standard features derived from a normalized co-occurrence matrix include energy, entropy, contrast, homogeneity and correlation features, among which the feature entropy measures the randomness of gray level distribution.

Entropy is defined as the measure of information content in a probability distribution. When processing images, the probability distribution is depicted by the histogram of intensity values. If probability of intensity is p_k and an image has l possible intensity values, the entropy can be calculated as:

$$H=-\sum_{k=1}^{l} p_k \log_2 p_k \tag{2}$$

Local entropy for processing image pixels can be calculated from formula (2). Local entropy of each pixel is computed by means of the intensity values of local neighborhood. And the neighborhood is specified beforehand. Entropy value of a 3×3 local neighborhood around each pixel array is returned as an array for non-text layer image. Meanwhile, local standard deviation is calculated as the standard deviation of a 3×3 local neighborhood. The outputs of both local entropy and local standard deviation of corresponding pixels are summed up to describe the intensity images containing all the non-textual objects. The thresholding is applied on the sum-up images of local texture entropy and local standard deviation to detect homogeneous connected components.

Aiming at the reflowable reconstruction of PDF documents, the purpose of the graphical composite segmentation lies in the rectangular bounding box of a holistic graphic composite which is mostly depicted by path operations in PDFs, rather than the continuous fine edge boundaries of the detailed contents of graphic. Hence, the morphological filter consisting conditional dilation is utilized to fill the holes, which can be formulated as:

$$Y_k=(Y_{k-1}\oplus D)\cap A^c, \qquad k=1,2,3,\cdots \tag{3}$$

where $Y_0=s$, and s is the start filling point belonging to non-boundary background points inside boundary. D is the symmetric structure element for dilation operation $\oplus$, which is conditioned at each step with A^c, the complement of A, and A is the set of connected boundary points of a region. The region filling process is stopped at iteration step k if $Y_k=Y_{k-1}$. It is applied on each detected homogeneous connected component.

After region filling process, the outside bounding box is then identified on the specific connected component. Up till now, the non-textual .png image I is segmented into n partitions R_n. Each subregion R_i, $i=1,2,\cdots,n$ is a connected component. In most cases, a whole graphic figure consists of multiple connected components. From the human vision perspective, further merging operations, even a few splitting operations, are needed to group CCs (connected components) into desired regions based on predefined criteria, which is closely related with the inter text line space. In this application context, our merging criterion is the proximity in spatial distance. For each bounding box of CCs, the distance along horizontal direction between any two boxes is defined as:

$$d_x=L(i,j)-R(i,j) \tag{4}$$

where $L(i,j)=\max(Box_i.left_x, Box_j.left_x)$, $R(i,j)=\min(Box_i.right_x, Box_j.right_x)$. The vertical distance is defined similarly with top and bottom of enclosing box. When two boxes are overlapped or attached to each other within the average inter text line space, the merging process will be performed, and the new bounding box is returned as graphic composites. For the purpose of simplicity, the decorating straight lines are removed. After the region growing process, the final graphic composite candidates can be identified for further integration with analysis results of text layer.

4.2 Analysis on Text based Layer

The goal of this section is to clarify whether the text element belongs to graphic composites. In this paper, two methods are introduced for text based layer analysis.

4.2.1 Clustering of text lines and text blocks

For all the text primitives provided by PDF parser, a bottom-up method can be used for text elements clustering. As is also explored in Ref. 5, text elements are first merged into text line, following by growing text lines into text blocks. This process is technically feasible in PDF documents, even with multi-columned ones[4,19]. Generally, page columns based on whitespace analysis approach[20] are detected in the first place, after which the text lines are clustered. For the advantages of digital born PDF documents, the positions of bounding boxes provided are gridded. In a manner of left to right and top-down processing order, two text elements are allowed to be clustered into a text line segment with the condition that these two text elements are adjacent in spatial distribution.

For text block identification, features including font size, inter text line spacing, left alignment and right alignment are taken into consideration. Within a document, the dominant font, also called as main font, is counted, which plays an important role in classifying the body texts and illustrative texts. Figure 4 (a) gave an example of the text block segmentation. As can be seen, all the text blocks within the body text region are successfully detected, but the illustrative text segment "Europas Arierien" is mistakenly identified as title text line.

Given the block segmentation results, each block enclosing box is compared with the segmented graphic composite candidate in non-text layer according to spatial distribution and font size information, integration is carried out when the illustrative text elements are attached to or overlapped with the segmented graphic composite, which is implemented by calculating both horizontal distance and vertical distance presented in Section 4.1.

4.2.2 Minimal spanning tree text clustering

From vision based perspective, page text elements (primitives) can also represent a connect component. Each text element correspond to a vertex in a two dimensional space, which is used for constructing graph. 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. In this application, the elements in V are the centroids of the bounding boxes extracted from PDF parser. All the text elements can be connected by establishing a neighborhood system by using Delaunay tessellation.

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. 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 vertices or primitives in a page, the spanning tree of the page has $n-1$ 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:

$$w(F)=\sum_{(v_i,v_j)\in F} w(v_i,v_j) \tag{5}$$

Minimal spanning tree requires that the sum of the edge weights is minimal among all other possible spanning trees of the same graph. In this paper, the weight for each edge $(v_i,v_j)\in E$ is selected as:

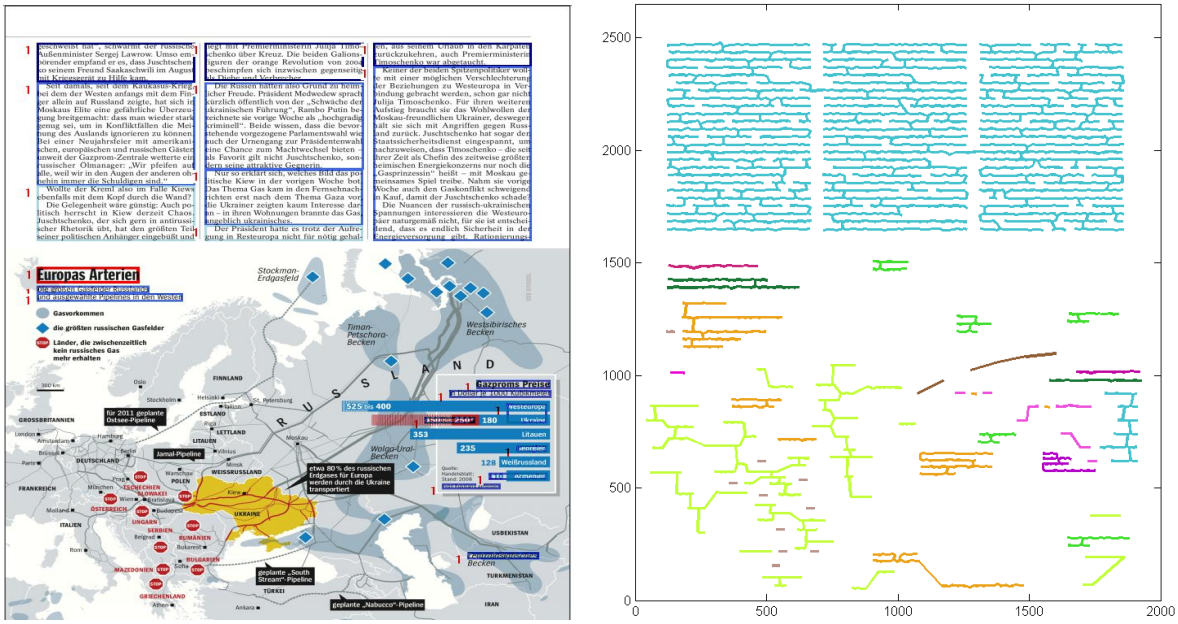
$$w(v_i,v_j)=f_E(v_i,v_j) \tag{6}$$

where $f_E(v_i,v_j)$ is the Euclidean distance function . A minima spanning tree of page graph is built by the Kruskal algorithm. To identify the feature difference of body texts and illustrative texts, font size information is taken into consideration for segmenting the graph tree formed by MST of each page elements into sub-tree by cutting unqualified edges connecting the CCs. The input is a page graph $G=(V,E_{MST})$ with n vertices. The output is a partition of V into a group of homogenous blocks $B=(B_1,\cdots,B_L)$. The algorithm procedure is as follows:

Step 1: Select an edge (v_i,v_j) from the MST E_{MST} iteratively.

Step 2: To check whether two vertices connecting the edge are in the same component, we use set-find operation on each vertex.

Step 3: According to the font size similarity of two vertices, set-union operation is used to merge two components. And a disjoint-set data structure is designed to contain the disjoint sets of elements. Each set $B=(B_1,\cdots,B_L)$ is a sub-tree from the current E_{MST}.



(a) (b)

Figure 4. (a) An example of text block identification using text line merging rules; (b) Text elements clustering based on partition of minimum spanning tree.

Figure 4 (b) gave an example of the text elements clustering based on minima spanning tree and disjoint-set based partition of MST edges. In similar way, given the text clustering groups, text body block is excluded from integration process according to the statistic distribution of edge weights and the number of the enclosed text elements, which is shown at the top block in Figure 4 (b). Each block box of illustrative texts candidate is compared with the segmented graphic composite candidate in non-text layer. And then integration process is carried out when the illustrative text blocks are attached to or overlapped with the segmented graphic composite.

5. RESULTS

To evaluate our proposed algorithm, three datasets are applied. A large variety in language, page layouts has been considered in the datasets. It is consisting of English with 264 pages, Chinese with 168 pages and German with 64 pages around the proportion of 4:3:1. The evaluated document pages in Chinese were selected from 20 e-books with subjects ranging from natural science to social science, which are provided by Founder Apabi digital library. The English pages were crawled from web, which covers conferences and journal papers, and both three columned English and Chinese documents are from modern magazines like "National Geographic", "Advances in Digital Computer" magazines. The German pages are from magazines like "Der Spiegel". The following Table.1 is given in detail for the configuration of datasets.

As is pointed out in 5, the precise quantitative evaluation for books and magazines requires ground truth. Although the construction of evaluation set is very time-consuming, preliminary work such as manual labeling platform has been already carried out, as is shown in the example of Figure 2(b), which will be further carried out in evaluation of both low-level and high-level page segmentation. In this paper, we measure the performance of accuracy by manually counting the number of correct segmented graphic composites, divided by the accurate number of graphics provided in our labeled ground truth. Correct segmented graphics need to consider both accuracy of the illustrative text elements and visually holistic graphic composites. Taken pages in Figure.5 as examples, there is one illustrative graphic composite in the bottom part and three column separating straight lines as human vision perceives in Figure.5 (a). As is discussed

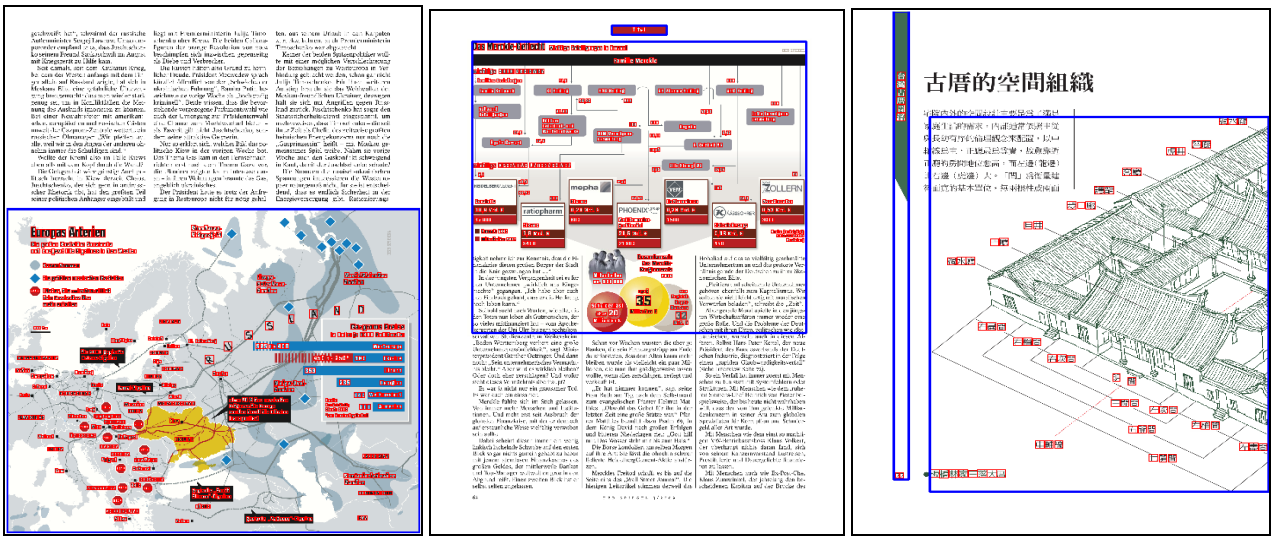
previously, the decorating straight lines serving as separating columns are removed. We can notice, with all the embedded text elements included, the map graphic composite is successfully detected, which indicates the accuracy is 100% for this page. Figure.5 (b) and (c) are tricky cases. The reason is that we use the bounding boxes to enclose each graphic composites for computational and representation conveniences. Figure.5 (b) is of T shape intersecting with text body region, which is quite different from the regular rectangular graphic composites without overlapping with text body. Using the proposed multi-layer method, the illustrative text elements are identified and merged into the T shape graphic composite, and four decorative lines are excluded for convenience. Similarly, Figure.5 (c) has two graphic objects according to our manual measure, without considering the decorative lines. The segmentation results gave us two graphic composites with text elements integrated accurately, and the accuracy is guaranteed as well. All the IDs appearing in xml file of illustrative text elements are saved as ID sets for graphic layout analysis, which are bounded with boxes in segmentation results of Figure 5.

For the testing on the 496 pages documents, the segmentation results of graphic composite integrating with texts are give in Table.1. The accuracy ranges from 42.8% to 100%. For general books, conference or journal pages, the segmentation accuracy are higher than that of magazines. Due to layout complexity and typesetting diversities in magazines, it is more difficult to segment graphics accurately into the desired region. In regard of processing time, the consumption is also higher for magazine document pages. For Chinese document pages, one-columned and tree columned pages only achieve over 77.8% and 42.8% accuracy mostly due to layouts complexity. In majority cases, the proximity of two types of graphics, such as an intensity image accidentally attached by decorating straight line, deteriorates the segmenting performance. For better handling of such cases in post processing, the accuracy can be greatly improved, but with the expenses of time complexity.

As can been seen, the average processing time for each PDF page with graphic composites ranges from 2.4 to 9.7 seconds on a personal computer (3GHz CPU, 3G RAM), which depends on the number of graphic composites within each page document and the processing of text elements clustering. The existence of large amount small graphical fragments, containing small dashes and pictorial letters can be very time consuming for segmenting performance.

Table 1. Quantitative Evaluation

Column	Language	Num. Pages	Num. Graphics	Num. Segmented Graphics	Accuracy	Processing Time (sec)
one column	English	97	144	117	81.3%	3.0
	Chinese	32	144	112	77.8%	9.1
	German	3	17	17	100%	3.8
two columns	English	128	190	178	91.5%	3.9
	Chinese	86	317	294	92.7%	2.4
	German	9	65	63	96.9%	3.8
three columns	English	39	103	79	76.7%	6.8
	Chinese	50	257	110	42.8%	9.7
	German	52	225	156	69.3%	5.4



(a) (b) (c)

Figure 5. Selected graphic composite segmentation results: (a) and (b) segmenting results of three columned pages from "Der Spiegel"; (c) segmenting result of one-columned Chinese page from the book "Taiwai Historical Architecture".

6. CONCLUSION

In this work, a multi-layer segmentation scheme is proposed to segment the graphic composites to cope with the complex layouts appeared in PDF documents. The inherent advantages of PDF like low level formatted inputs are exploited for layout analysis. Hence, the bounding boxes of the page elements embedded by PDF are exported and converted from the metric of logic units to pixels. The contributions of our work mainly focus on following aspects:

- Analysis on non-text and text based layer are carried out for physical structure layout extraction, respectively.
- Image-based connected components extraction is incorporated into PDF document analysis to replace trivial clustering of numerous path operations for graphic segmentation.
- Text elements clustering based on MST and disjoint-set is proposed for identifying illustrative texts.
- Integration of text layer and non-text layer analysis results is applied for holistic graphic segmentation.

The experimental results on PDF books and magazines gave us more confidence on multi-layer segmentation method when compared with path primitives clustering based graphic segmentation in terms of handling PDF documents with various layouts.

ACKNOWLEDGEMENTS

This work was supported by the National Basic Research Program of China (973 Program) (No.2012CB724108).

REFERENCES

[1] Simone, M., Emanuele, M., "Table of contents recognition for converting PDF documents in e-book formats," in 10th ACM symposium on Document, 73-76 (2010).
[2] Marinai, S., Marino, E., Soda, G., "Conversion of PDF Books in ePub Format", in 2011 International Conference on Document Analysis and Recognition, 478-482 (2011).
[3] Hadjar, K., Rigamonti, M., Lalanne, D., Ingold, R.,, "Xed: a new tool for extracting hidden structures from electronic documents," in Proceedings of First International Workshop on Document Image Analysis for Libraries, 212- 224 (2004).

[4] 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, 78740U 78741-78749 (2011).
[5] Lin, X., "Header and footer extraction by page-association", in Proc. SPIE Conference on Document Recognition and Retrieval X, 164-171 (2003).
[6] Déjean, H., Meunier, J., L., "A system for converting PDF documents into structured XML format", in Proc. DAS' 06, 129-140 (2006).
[7] Gao, L., Tang, Z., Qiu, R., "A mixed approach to auto-detection of page body," in SPIE-IS&T International Conference of Document Recognition and Retrieval XV, (2008).
[8] Fang, J., Gao, L., Bai, K., Qiu, R., Tao, X., Tang, Z., "A Table detection method for multipage PDF documents via visual seperators and tabular structures," in International Conference on Document Analysis and Recognition, 779-783 (2011).
[9] Lin, X., Gao, L., Tang, Z., Lin, X., Hu, X., "Mathematical formula identification in PDF documents," 2011 International Conference on Document Analysis and Recognition, 1419-1423 (2011).
[10] Tombre, K., "Graphics recognition: the last ten years and the next ten years," Lecture Notes in Computer Science 3926, 422-426 (2006).
[11] Fan, J., "Text segmentation of consumer magazines in PDF format," in International Conference on Document Analysis and Recognition, 794-798 (2011).
[12] Antonacopoulos, A., Pletschacher, S., Bridson, D., Papadopoulos, C., "ICDAR2009 page segmentation competition," in International Conference on Document Analysis and Recognition, 1370-1374 (2009).
[13] Chowdhury, S., P., Mandal, S., Das, A., K., Chanda, B., "Segmentation of text and graphics from document images", in 9th International Conference on Document Analysis and Recognition, 619-623 (2007).
[14] Fletcher, L., A., Kasturi, R., "A robust algorithm for text string separation from mixed text/graphics images," IEEE Transactions on Pattern Analysis and Machine Intelligence 10(6), 910-918 (1998).
[15] Ahmed, S., Weber, M., Liwicki, M., Dengel, A., "Text/Graphics segmentation in architectural floor plans," 2011 International Conference on Document Analysis and Recognition, 734-738 (2011).
[16] Chao, H., "Graphics extraction in PDF document," in Document Recognition and Retrieval X, Santa Clara, CA, USA, pp, 2003.
[17] Shao, M., Futrelle, R., P., "Recognition and classification of figures in PDF documents," in Graphics Recognition. Ten Years Review and Future Perspectives. LNCS, 239-251 (2006).
[18] Fang, J., Tao, X., Tang, Z., Qiu, R., Liu, Y., "Dataset, ground-truth and performance metrics for table detection evaluation," 2012 10th IAPR International Workshop on Document Analysis Systems, 445-449 (2012).
[19] Bloechle, J-L., Lalanne, D., Ingold, R., "Ocd: an optimized and canonical document format," 2009 International Conference on Document Analysis and Recognition, 236-240 (2009).
[20] Breuel, T., M., "Two geometric algorithms for layout analysis," Document Analysis Systems (DAS'02), 188-199 (2002).