Review of chart recognition in document images

Conference Paper *in* Proceedings of SPIE - The International Society for Optical Engineering · February 2013

DOI: 10.1117/12.2008467

Xiaoqing Lyu
Peking University

Review of Chart Recognition in Document Images

Yan Liu*[a,b,c], Xiaoqing Lu[a], Yeyang Qin[a], Zhi Tang[a], Jianbo Xu[b]
[a]Peking University, Zhongguancun North Street, Beijing, Chiana; [b]State Key Laboratory of Digital Publishing Technology (Peking University Founder Group Co. Ltd.), Chengfu Street, Beijing, China; [c]Postdoctoral Workstation of the Zhongguancun Haidian Science Park, Beijing, China.

ABSTRACT

As an effective information transmitting way, chart is widely used to represent scientific statistics datum in books, research papers, newspapers etc. Though textual information is still the major source of data, there has been an increasing trend of introducing graphs, pictures, and figures into the information pool. Text recognition techniques for documents have been accomplished using optical character recognition (OCR) software. Chart recognition techniques as a necessary supplement of OCR for document images are still an unsolved problem due to the great subjectiveness and variety of charts styles. This paper reviews the development process of chart recognition techniques in the past decades and presents the focuses of current researches. The whole process of chart recognition is presented systematically, which mainly includes three parts: chart segmentation, chart classification, and chart Interpretation. In each part, the latest research work is introduced. In the last, the paper concludes with a summary and promising future research direction.

Keywords: Chart segmentation, Chart classification, Chart interpretation

1. INTRODUCTION

Chart was a powerful tool for plotting statistical data in a visual way, which made the document contents more comprehensive and informative, thus, it was suggested in many research fields [1-4]. With the rapid increase of document literatures, methods for documents analysis had been studying by lots of researchers [5-7] in the past decades and covered a wide range of image types, from journal papers to engineering drawings, from regular documents analysis to affine transform invariance. Research activities in the field of document image analysis can be mainly classified into two categories, that is, the text processing and non-text processing e.g. figure, graphics, and diagram. Text recognition from document images had been accomplished simply by lots of optical character recognition (OCR) systems [8-9]. Traditional OCR system focused on segmenting and recognizing the text from input document images. The output of traditional OCR system mainly included two parts: the recognized text in the electronic form and the layout information of the original document [10]. Thus, many chart segmentation methods were implemented using the text-labeling mask generated by OCR. For the non-text objects in the document, such as chart, illustration, photos, the OCR system simply segments them out from the original document page and outputs them as images. Thus, the OCR system cannot interpret the contents in charts. In addition, for the some cases, the OCR system cannot distinguish whether the text belongs to a chart or the main body of the document.

Publications related to OCR recognition of text were in abundant [11-12], and the success of these algorithms depended on successful separation of text and graphical part in document images. As the precondition of chart recognition, the task of separating graphics from text was considered important since they contained important information that cannot be recognized efficiently by the existing OCR system [13]. Thus, it should be a good supplement to add chart recognition component to the OCR system. However, as the conclusion of Cushman [14], the correct recognition of OCR system should be 98% at least to make it competitive, while most current chart recognition system cannot meet this comparable performance standard. There lacked of efficient methods to correctly recognize and interpret the figures and drawings in complex documents. Even though chart recognition had been studied since 1990s, there was little research work and practical results reported on recognizing and interpreting chart images, comparing to those on other types of document images such as forms, tables or engineering drawings [15-18].

On the basis of previous research, we concluded the process of chart recognition into three parts: chart segmentation, chart classification, and chart interpretation respectively. As shown in Fig.1, the function of chart segmentation was to separate the non-text from the text in document images, where OCR was a tool relied on generally. Chart classification

*lyan2006@gmail.com;

Visualization and Data Analysis 2013, edited by Pak Chung Wong, David L. Kao, Ming C. Hao, Chaomei Chen,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 8654, 865410 · © 2013 SPIE-IS&T
CCC code: 0277-786X/13/$18 · doi: 10.1117/12.2008467

technique classified coarsely the non-text part into three categories: illustration, photo, and chart. Only the chart objects were classified further for a specified chart style. This part attracted the most attention, thus, many existing methods could be referred [19-21]. The part of chart interpretation belonged to the high-level chart recognition, where the data included in each kind of chart was interpreted and the chart can be redesigned according to the extracted if necessary. The organization of this paper also followed the process of chart recognition and introduced the three parts in order.

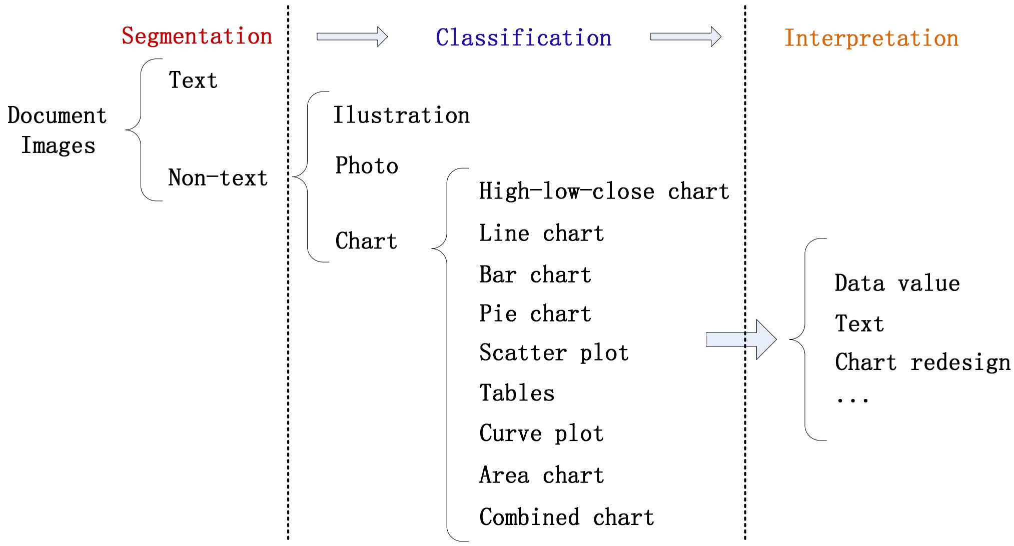
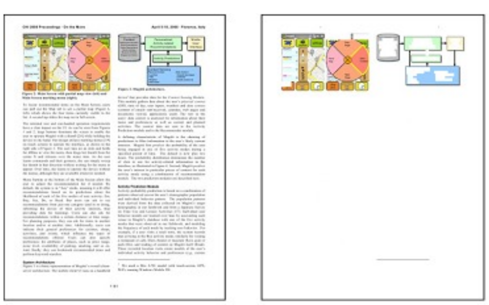


Figure. 1 General process of chart recognition

2. CHART SEGMENTATION

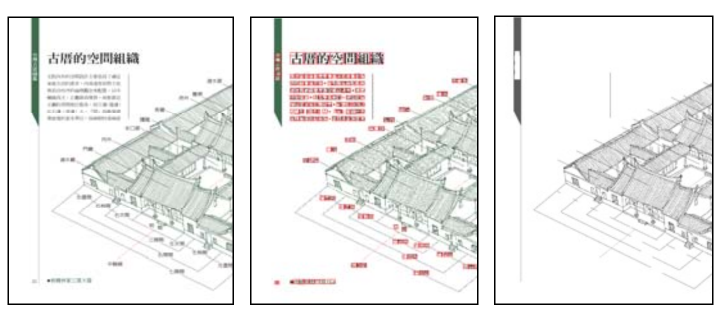
Efficient chart segmentation was a necessary precondition for chart recognition, even though it was still a hard challenge to variety of documents. Many studies in image key-point feature describing[22-24] had proved that segmentation sometimes can be avoided, but there was no literature show its feasibility for chart segmentation in document images. Therefore, this paper only summarized the segmentation based chart recognition techniques.

Few papers presented the whole chart recognition process explicitly, because separating text and non-text was still a hard problem to solve. Many chart recognition methods accomplished text segmentation with OCR techniques[25], although which was not very efficient to some complex documents as analyzed above. Fig.2 showed the process of text and non-text segmentation process with OCR, where the (b) was the text/non-text mask for segmentation. In his paper, it was not advised to use OCR system only for text and non-text separation, because some annotation texts surrounding the chart may be misclassified. Similar text segmentation problem was shown in Fig.3 where the text was labeled by the document XML description. The text belonging to the illustration of buildings was removed blindly, as was greatly different with the analysis in Ref. [26] and motivated the study of pure image based text/non-text segmentation method.



(a) (b)

Figure. 2 Text and non-text parts segmentation. (a) Example of a document page image. (b) The masked image created from using the OCR text bounding boxes.



(a) (b) (c)

Figure. 3 Text segmentation by the document XML description. (a) Initial document image. (b) XML labeled text by red rectangle. (c) Separate text from non-text

Approaches focused on separating graphics from text could be grouped into three categories. They were directional morphological filter[27], extraction of lines and arcs[28], and connected component analysis[29, 30]. The choosing of different methods was relied on the system models: top-down or bottom-up[8][31]. For the top-down approach, a basic model was derived and used to search for mapping in the given image. This approach was often based on run length smoothing (RLS) algorithm, and split the document into paragraph blocks, into text lines[32-33]. The RLS algorithm was applied in both horizontal and vertical directions to binary image, and then the logical AND or OR operation of these two output images would consist of blocks which represented different functional areas. Such blocks can be further classified and assigned to a specific class. The algorithm first viewed the entire image globally and then divided the image into small areas by x-y cuts[34]. For the bottom-up approach, some basic features were extracted and used for constructing an existing model. This approach typically involved grouping pixels as connected components which were then merged into successively larger regions. Jain & Bhattercharjee[35] described a relevant segmentation method based on multichannel filtering technique using Gabor filters. J.S.Payne et al[8] proposed a novel texture recognition strategy, which involved the assembly of nth order co-occurrence information within a processing window. The resulting spectrum was processed by nearest neighbor classifier, and discontinuities in the classification as the operating window scanning the image were recognized as texture boundaries. In their research, this technique was applied for the first time to segmenting newsprint in order to locate the text regions for effective data capture. Fig.4 showed a generic document image segmentation process. In order to separate text from graphics, connected components were constructed by grouping pixels with similar intensities and were eight neighbors of each other[13,36]. A series of filters were applied to classify the connected components into textual components and graphical components. These filters taken into consideration the area (number of pixels), the height and width, the height/width ratio and the black pixel (assuming a black image on white background) density of each connected component constructed. Thresholds that were obtained through training examples were used for separating text from non-text regions.

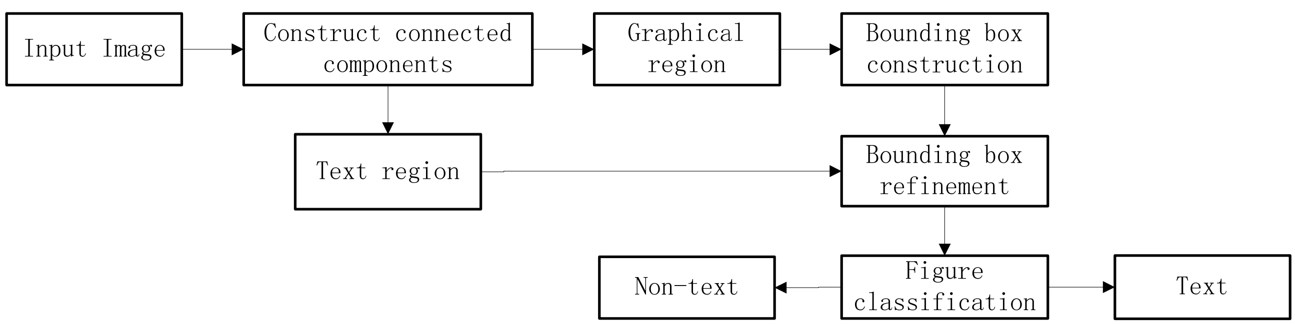
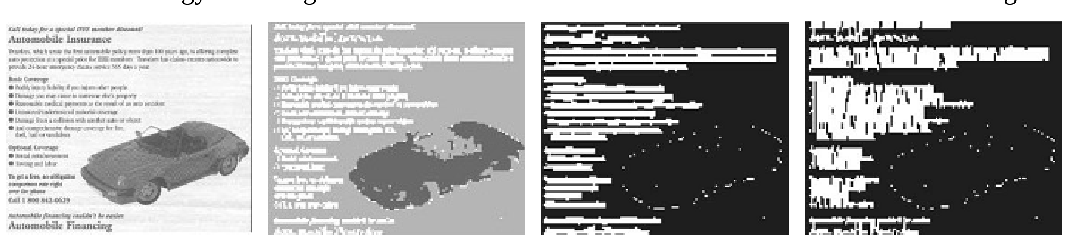


Figure. 4 One method for text and non-text regions separation

Since text regions in a document were quite different from the texture of background or not-text regions, thus, it was possible to easily separate text from the non-text regions, such charts, illustration and photos. Document segmentation can be implemented in a number of ways. In Ref.[37], a multiple scale approach was introduced and the performance of wavelet packet was investigated. And as the conclusion of [37], multi-resolution properties of wavelet transform were useful for applications such as segmentation, classification, and texture discrimination. In Ref.[38], a criterion utilizing ratio of the mean energy in the four low-frequency channels to that in three middle-frequency channels was suggested to distinguish smooth images from textured images. Since picture images were different from the text in the continuity of their distributions, and the wavelet coefficients in the high-frequency bands usually followed a Laplacian distribution[39], features can be defined depending on the shape of the histogram of wavelet coefficients. In Ref.[33], an efficient document segmentation method based on cubic B-spline wavelet was proposed, where the structural features: standard deviation and energy of the region are defined for classification. The test results are listed in Fig.5.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure. 5 Performance of the proposed segmentation method on a document image. In this image, black, white and gray colors label the picture, text and background, respectively: (a) original image; (b) output image of three-means segmentation; (c) horizontal merging; (d) vertical merging; (e) OR operation on (c) and (d); (f) median filtering result; (g) dilation processing; (h) final segmentation result.

3. CHART CLASSIFICATION

After the chart segmentation, text from non-text regions had been separated. In the following, it is necessary to classify the chart from the non-text regions. There had been lots of publications to present charts classification methods[7,21,40]. In the paper, we classified the non-text regions into three categories coarsely: chart, illustration and photo, as illustrated in Fig.6. This kind of classification was similar to the first level classifier output in Ref. [41]. For non-text regions classification, image features can be extracted to distinguish the classes. Texture feature[42] provided a better description of the selected regions and was widely used in chart classification, where the gray level co-occurrence matrix (GLCM) features played an vital role in many texture based image analysis. As proposed in Ref. [7], the area, median, minimum and maximum intensity, contrast, homogeneity, energy, entropy, mean, variance, standard deviation and correlation features had to be considered.

(a) Pie chart (b) Bar chart (c) Flow chart (d) Illustration (e) Photo

Figure. 6 Coarse classification of non-text parts

When excluding illustrations and photos from non-text regions, the remained charts were classified further. Nowadays, there still lack of classification standard. The definition for chart is still blank. In this paper, we named the figures used for statistic data illustration as charts whose content can be redesigned. In Fig.7, twelve styles of chart were illustrated where the first eleven graphs were 2D charts. The last one was 3D chart where only one pie chart was illustrated as an example. Each kind of illustrated chart also included lots of mutations which made the chart recognition more complex. Most of the existing studies focused on only one or several kinds of the charts.

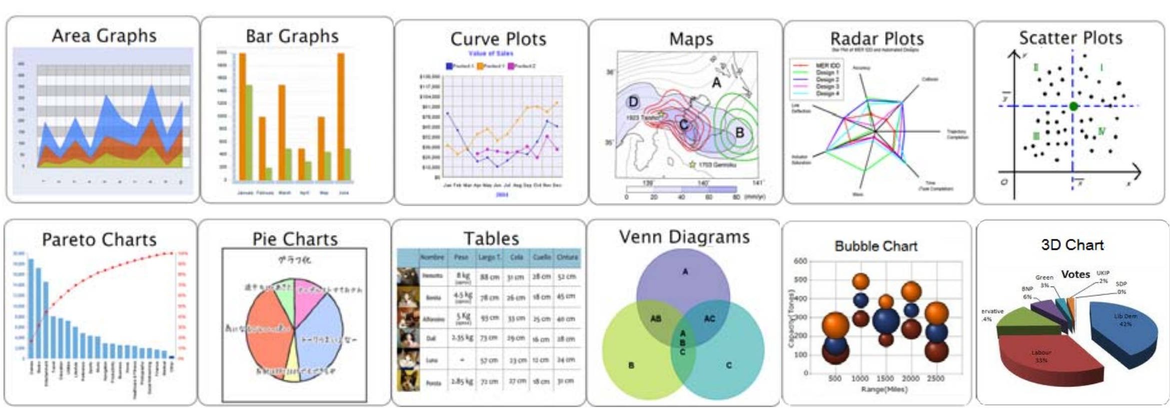


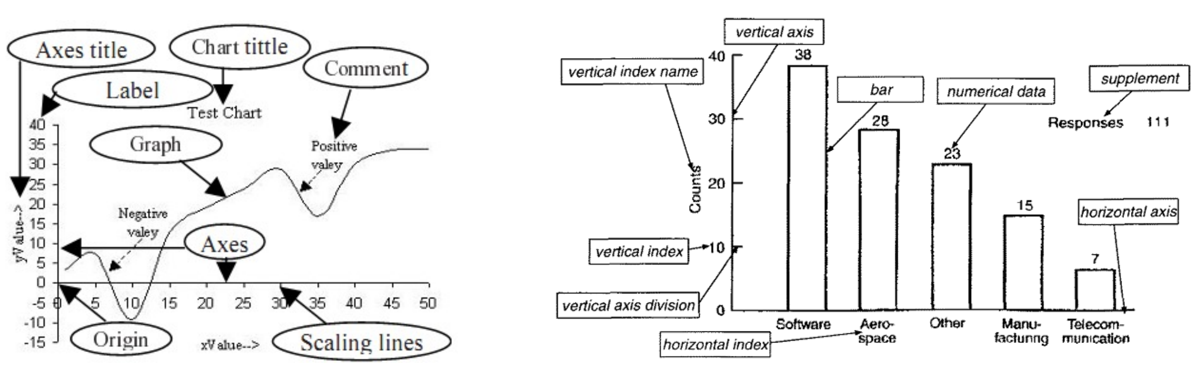
Figure. 7 Fine classification of chart styles

For robust detection to different charts, researcher introduced some high-level textual and semantic factors and some predefined primitives to chart's structure. Zhou and Tan[43] proposed Hough-based bar chart detection and segmentation which combined with the syntactic analysis. In their follow-up work[44], a modified probabilistic Hough transform algorithm was developed to speed up the standard Hough transformation technique to recognize the bar chart and robust to hand-drawn bar chart. In their latest works[15], two kinds of chart recognition methods: hidden Markov model based and neural network based method were proposed. Both of the methods were tested in four kinds of chart: high-low-close chart, line chart, bar chart, and column chart. Relying on multi-layer feed-forward neural network training, the mean precision was about 90%. Other recognition techniques like curvature estimation[45], vector-based technique[46] and raster-to-vector conversion algorithm[47] were proposed to identify line chart and pie chart. Futrelle et al[48] and Shao et al[49] proposed a scheme for recognizing and classifying vector format graphics in PDF documents using spatial analysis and classified charts into five categories: line, bar, curve, tree and other charts. In the research of Karthikeyani et al[7], eleven types of texture features and three classifier, namely multilayer perception, support vector machine and K nearest neighbor, were used. Eight types of charts were considered, namely, 2D, 3D bar chart, 2D, 3D pie chart, 2D, 3D doughnut chart, line chart and mixed chart. Yokokura et al[50] proposed a schema-based framework to graphically describe the layout relationship of the bar chart based on vertical and horizontal projection. Although project method was a generally used technique, its application for different chart styles was restricted due to its simplicity.

4. CHART INTERPRETATION

In order to make the chart, consisting of graphical and textual components, computer readable and redesign able, it was usually to construct chart model[51] . Since lack of standard for chart style, each kind of chart had to be modeled separately (widely used but not praised). As shown in Fig.8 whose primitives, such as title, axes names, legends, etc, were labeled with semantic annotation. Both the textual information and graphical information were essential to be associated for redesigning. However, as summarized in Ref. [52], the primitives appeared in each kind of chart styles might be greatly different. Thus, the study on of chart redesigning ignored some unnecessary primitives and extracted only some essential information such as vertical and horizontal indexes, and the legend.

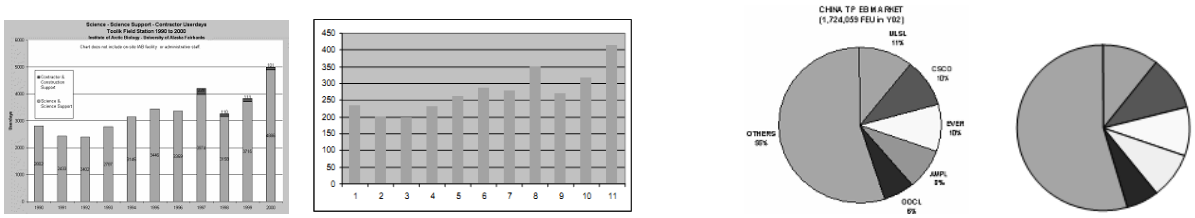
Fig.9 showed two kinds of chart redesigning results, where (a) was the bar chart redesign and (b) was the output: pie chart redesigned. Both of the redesigned results were with no text annotation[31] and the output styles were not changed. Some research improved the former and redesigned the chart with both the text and graphics[52]. Fig.10 showed another two relevant results, where in (a) the bar chart was redesigned to the table[51] and in (b) the pie chart was redesigned to the bar chart[2]. Both of the redesigned results were with text annotation. In addition to 2D chart recognition, 3D chart recognition and redesign were being studied by lots of researchers[51][53]. Charts with different resolution, edge distortion, horizontal shearing, motion blur and noises were being studied[10]. And in the study of Yanping Zhou[45], hand-drawn charts were redesigned accurately using modified probabilistic Hough transformation. Although, lots of wok had been done on chart interpretation, most of the methods could be used only for a special or several kinds of chart images. It is still lack of efficient method to solve multiple kinds of chart in extensive documents. For example, it was not easy to determine whether the primitive or annotation belonged to the chart or not for flow chart and combined complex charts. In addition, few works talked about the robustness, such as color invariance, resolution invariance and even distortion invariance, thus human prior knowledge might be an important cue.



(a) Primitives of line chart

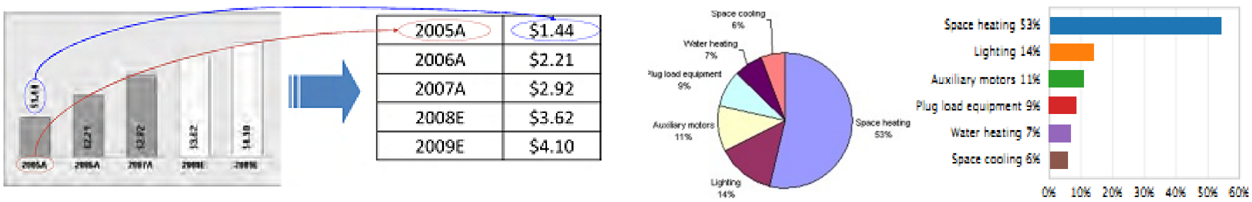
(b) Primitives of bar chart

Figure. 8 Constructed chart model



(a) Bar chart redesign (b) Pie chart redesign

Figure. 9 Chart redesign with no annotation



(a) Bar chart redesign (b) Pie chart redesign

Figure. 10 Chart redesign with annotations

5. CONCLUSIONS

Charts were widely used in scientific, financial and other analytical documents as a graphical representation of numerical and qualitative data. Chart recognition technique was a good supplement of OCR system for document analysis and data retrieval. However, the existing chart recognition systems usually were used in laboratory. Only a few of them were applied in special kind of chart recognition for company. The major challenges hindering the development of char recognition included six aspects: (i) the great diversity of chart types and styles, which were somewhat very subjective without standard templates. (ii) the flexibilities in the structural arrangement. For the same statistical data, different researchers might give great different structural curve chart design. (iii) the difficulties of adequately describing the syntax and semantics of the complex charts. Although text can be recognized by OCR, the role of the text in chart was not even very clearly defined and formulated. (iv) the difficulties in handling degraded, distorted, rotated, scale-different or noisy input, even hand-drawn input. (v) the lack of generality of chart recognition methods. Most of researches were concerned on one or several solid chart styles. There lacked a systematic regulations to deal with multi-kinds of general chart recognition. (vi) the lack of ground truth dataset for test and comparison between different methods. Even though the development of chart recognition methods had been puzzled by these problems listed above, while with the developing of the document analysis, chart recognition technique as a necessary supplement to existing OCR system would attract more and more attention.

ACKNOWLEDGMENTS

This work is supported by the National Basis Research Program of China (No.2010CB735908), Beijng Postdoctoral Scientific Research Funding and China Postdoctoral Science Foundation.

REFERENCES

[1] B.Twersky, J.Zacks, P.Lee and J.Heiser, “Lines, blobs, crosses and arrows: diagrammatic communication with schematic figures,” Theory and Application of Diagrams, 221-230(2000).

[2] Manolis Savva, Nicholas Kong, Arti Chhajta, Li Fei-Fei, et al., “Revision: Automated Classification, Analysis and Redesign of Chart Images,” Proc. of the 24th ACM Symposium on User Interface Software and Technology, 393-402(2011).

[3] Sudhindra Shukla, Ashok Samal, “Recognition and Quality Assessment of Data Chart in Mixed-mode Documents,” International Journal on Document Analysis and Recognition 11(3), 222-126(2008).
[4] Daehyun Kim, Balaji Polepalli Ramesh, Hong Yu, “Automatic Figure Classification in Bioscience Literature,” Journal of Biomedical Informatics, 848-858(2011).
[5] slam,R. Saha, R.S, and Hossain, A.R., “Automatic Reading from Bangla PDF Document Using Rule Based Concatenative Synthesis,” International Conference on Signal Processing Systems, 521-525(2009).
[6] Martinez Alvarez, R.P., Costas Rodriguez, S., Gonzalez Castao, F.J. and Gil Castieira, F., “Automated Document Conversion System for Simple Multimedia Platform,” The 7th IIEEE Consumer Communications and Networking Conference, 1-2(2010).
[7] V.Karthikeyani, S.Nagarajan, “Machine Learning Classification Algorithms to Recognize Chart Types in Portable Document Format (PDF) Files,” International Journal of Computer Applications 39(2),1-5(2012).
[8] J.S.Payne, T.J. Stoham, and D.Patel, “Document Segmentation Using Texture Analysis,” Proc. of the 12th IAPR International Conference on Computer Vision & Image Processing, 380-382(1994).
[9] Takuma Yamaguchi, and Minoru Maruyama, “Feature Extraction for Document Image Segmentation by pLSA Model,” The 8th IAPR Workshop on Document Analysis System, 53-60(2008).
[10] Weihua Huang, “Scientific Chart Image Recognition and Interpretation,” PHD paper of National University of Singapore, 2008.
[11] Chellapilla,K.,Simard,P., and Nickolov, R., “Fast Optical Character Recognition Through Glyph Hashing for Document Conversion,” Proc. of the 8th International Conference on Document Analysis And Recognition, 829-833(2005).
[12] IsIam, R., Saha, R.S. and Hossain, A.r., “Automatic Reading from Bangla PDF Document Using Rule Based Concatenative Synthesis,” International Conference on Signal Processing System, 521-525(2009).
[13] Dr.V.Karthikeyani, S.Nagarajan, “Scientific Chart Image Property Identification by Connected Component Labeling in PDF Files,” The 3rd International Conference on Electronics Computer Technology, 209-212(2011).
[14] W.cushman, P.Ojha, and C.Daniels, “Usable OCR: What are the Minimum Performance Requirements,” Proc. ACM SIGCHI Conference on Human Factors in Computing Systems, 145-151(1990).
[15] Yanping Zhou, Chew Lim Tan, “Learning-based Scientific Chart Recognition,” The 4th IAPR International Workshop on Graphics Recognition, GREC, 482-492(2001).
[16] Henry S.Baird, Siyuan Chen, “Pixel Accurate Document Image Content Extraction,” Proc. of The ACM Symposium on Applied Computing, 245-251(2011).
[17] M.r.Casey and H.S.Baird, “Toward Versatile Document Analysis Systems,” Proc. of the 7th IAPRA Document Analysis Workshop, 280-290(2006).
[18] M.A.Moll and H.S.Baird, “Document Content Inventory & Retrieval,” ICDAR, 93-97(2007).
[19] Urszula Markowska Kaczmar, Pawel Minda, Krzysztof Ociepa, “Toward Automatic Image Annotation Supporting Document Understanding,” Hybrid Artificial Intelligent Systems Lecture Notes in Computer Science, 420-427(2011).
[20] Md. Khademul islam Molla, kamrul Hasan Talukder, and Md.Altab Hossain, “Line Chart Recognition and Data Extraction Technique,” Intelligent Data Engineering And Automated Learning-Lecture Notes in Computer Science, 865-870(2003).
[21] Daehyun Kim, Hong Yu, “Hierarchical Image Classification in the Bioscience Literature,” Journal of Biomedical Informatics 4(5), 848-858(2011).
[22] Edward Rosten, Reid Porter, Tom Drummond, “Faster and Better: A Machine Learning Approach to Corner Detection,” PAMI, 105-119(2010).
[23] Ethan Rublee, Vincent Rabaud, Kurt Konolige, Gary Bradski, “ORB: An Efficient Alternative to SIFT OR SURF,” ICCV, 2564-2571(2011).
[24] David G.Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” IJCV, 91-110(2004).
[25] Dorothea Blostein, Edward Lank, Richard Zanibbi, “Treatment of Diagrams in Document Image Analysis,” The 1st International Conference on Theory and Application of Diagrams-Lecture Notes in Computer Science, 7-42(2000).
[26] J.M.Gloger, “Use of Hough Transform to Separate Merged Text/Graphics in Forms,” ICPR, 268-271(1992).
[27] Luo, H. and Kasturi, R., “Improved Directional Morphological Operations for Separation of Characters from Mpas/Graphics,” The Second International Workshop on Graphics Recognition, Algorithms and Systems, 35-47(1998).
[28] Kacem, A., Belaid, A. and hamed, M.B., “Automatic Extraction of Printed Mathematical Formulas Using Fuzzy Logic and Propagation of Context,” International Journal on Document Analysis and Recognition 4(2),97-108(2001).

[29] Fletcher, L. and Kasturi, R., “A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images,” PAMI 10(6),910-918(1998).

[30] Dr.V.Karthikeyani, S.Nagarajan, “Scientific Chart Image Property Identification By Connected Component Labeling in PDF Files,” The 3rd International Conference on Electronics Computer Technology, 209-212(2011).

[31] Weihua Huang, Chew Lim Tan, and Wee Kheng Leow, “Model-based Chart Image Recognition,” Lecture Notes in Computer Science, Graphics Recognition, Recent Advances and Perspectives, 87-99(2004).

[32] Frank Y.Shih, Shy-Shyan Chen, D.C.Douglas Hung, and Peter A.N., “A Document Segmentation, Classification and Recognition System,” Proc. of the 2rd International Conference on Systems Integration, 258-267(1992).

[33] Shulan Deng, Shahram Latifi, Emma Regentova, “Document Segmentation Using Polynomial Spline Wavelets,” Pattern Recognition, 2533-2545(2001).

[34] G.Nagy, S.C.Seth, S.D.Stoddard, “Document Analysis with an Expert System,” Proc. of ACM Conference on Document Processing Systems, 169-176(1988).

[35] Jain & Bhattachrjee, “Text Segmentation Using Gabor Filters for Automatic Document Processing,” Machine Vision and Applications, 169-184(1992).

[36] K.Tombre, S.Tabbone, L.Pelissier, B.Lamiroy, and P.Dosch, “Text/Graphics Separation Revisited,” DAS, 200-211(2002).

[37] A.Laine, F.Jian, “Texture Classification by Wavelet packet Signitures,” PAMI 15(11),1186-1191(1993).

[38] R.Porter, N.Canagarajah, “A Robust Automatic Clustering Scheme for Image Segmentation Using Wavelets,” IEEE Transaction on Image Processing 5(4), 662.-665(1996).

[39] Jia Li, R.M.Gray, “Text and Picture Segmentation by The Distribution Analysis of Wavelet Coefficients,” International Conference on Image Processing, 790-794(1998).

[40] Rafkind B, Lee M, Chang S, Yu H., “Exploring Text and Image Features to Classify Image in Bioscience Literature,” Proc. of the BilNLP Workshop on Linking Natural Language Processing and Biology HLT-NAACL 06, 73-80(2006).

[41] Urszula Markowska-Kaczmar, Pawl Minda, Krzysztof Ociepa, “Towards Automatic Image Annotation Supporting Document Understanding,” Lecture Notes in Computer Science, Hybrid Artificial Intelligent Systems, 420-427(2011).

[42] Haralic, R.M., Shanmugam, K. and Dinstein,I., “Textual Features for Image Classification,” IEEE Transactions on Systems, Man and Cybernetics 3(6),610-621(1973).

[43] Yanping Zhou, ChewLim Tan, “Bar Charts Recognition Using Hough Based Syntactic Segmentation,” Lecture Notes in Computer Science, Theory and Application of Diagrams, 494-497(2000).

[44] Yanping Zhou, ChewLim Tan, “Hough Technique for Bar Chart Detection and Recognition in Document Images,” International Conference on Image Processing, 605-608(2000).

[45] Rosin, Paul, “Segmentation of Edges into Lines and Arcs,” Image and Vision Computing 7(2):109-114(1989).

[46] Dori, D., “Vector-based Arc Segmentation in the Machine Drawing Understanding System Environment,” PAMI 17(11):1057-1068(1995).

[47] Song,J., Su, F., Chen, J., Tai, C.L. and Cai, S., “Line Net Global Vectorization: An Algorithm and Its Performance Analysis,” IEEE Conference on Computer Vision and Pattern Recognition, 383-388(2000).

[48] Robert P. Futrelle, Mingyan Shao, Chris Cieslik, and Andrea Elaina Grimes, “Extraction, Layout Analysis and Classification of Diagrams in PDF Documents,” ICDAR, 1007-1013(2003).

[49] Mingyan Shao, Robert P.Futrelle, “Recognition and Classification of Figures in PDF Documents,” Graphics Recognition-Lecture Notes in Computer Science, 231-242(2006).

[50] N.Yokokura and T.Watanabe, “Layout-based Approach for Extracting Constructive Elements of Bar-charts,” Graphics Recognition: Algorithms and Systems, 163-174(1997).

[51] Ales Mishchenko, Natalia Vassilieva, “Chart Image Understanding and Numerical Data Extraction,” The 6th International Conference on Digital Information Management, 115-120(2011).

[52] Sudhindra Shukla, Ashok Samal, “Recognition and Quality Assessment of Data Charts in Mixed-mode Documents,” International Journal on Document Analysis and Recognition 11(3),111-126(2008).

[53] Weihua Huang, Chew Lim Tan, and Wee Kheng Leow, “Elliptic Arc Vectorization for 3D Pie Chart Recognition,” ICIP, 2889-2892(2004).