(1) Name of the workshop:

Workshop on Multimedia Technologies

(2) Title of the paper:

The Extraction of Text/Graphs from Degraded Documents

(3) Short abstract:

This paper presents a simple and efficient algorithm to clarify the noisy background from a severely degraded document. This method can be applied to enhance the old dated historical documents such that the text to be clear readable condition and the photograph images to be preserved if there is any. We analyze the histogram of the document to find out the characteristics of text-, graph-, and background-pixels. By doing so we identify the graph-pixels and preserve their original gray values and apply Agent Growing Method [1] to the rest of the document in order to clear the noisy background. Our method shows a satisfying result in comparing with several known algorithms.

(4) Name, current affiliation, postal address, e-mail address, telephone number, and fax number for each author:

Shwu-Huey Yen[1](**顏淑惠**), Yi-Jin Chen[2](**陳羿瑾**), Mei-Fen Chen[3](**陳梅芬**)
Department of Computer Science and Information Engineering, Tamkang University
Tamsui, Taiwan, R.O.C.
Tel: 02-2621-5656 ext 2748, Fax: 02-2620-9749
e-mails:1. shyen@cs.tku.edu.tw; 2. 689190238@s89.tku.edu.tw; 3. 690190243@s90.tku.edu.tw

(5) Name of the contact author:

Shwu-Huey Yen

(6) Keywords:

degraded documents, extraction, histogram smoothing, histogram partition, agent-growing, morphological opening.

The Extraction of Text/Graphs from Degraded Documents

Shwu-Huey Yen[1](顏淑惠), Yi-Jin Chen[2](陳羿瑾), Mei-Fen Chen[3](陳梅芬)
Department of Computer Science and Information Engineering, Tamkang University
Tamsui, Taiwan, R.O.C.
(e-mail[1]: shyen@cs.tku.edu.tw, e-mail[2]: 689190238@s89.tku.edu.tw, e-mail[3]: 690190243@s90.tku.edu.tw)

ABSTRACT

This paper presents a simple and efficient algorithm to clarify the noisy background from a severely degraded document. This method can be applied to enhance the old dated historical documents such that the text to be clear readable condition and the photograph images to be preserved if there is any. We analyze the histogram of the document to find out the characteristics of text-, graph-, and background-pixels. By doing so we identify the graph-pixels and preserve their original gray values and apply Agent Growing Method [1] to the rest of the document in order to clear the noisy background. Our method shows a satisfying result in comparing with several known algorithms.

Keywords: degraded documents, extraction, histogram smoothing, histogram partition, agent-growing, morphological opening.

1. INTRODUCTION

For purpose of handling huge amount of historical degraded documents including dated books, newspaper, magazines, as well as those kept in the form of copies, photos or microfilms, etc., it is very important to reconstruct the original clean and readable conditions of these documents. Because most documents are mixtures of texts and graphs, a preprocessing of text/graph separation is by all means necessary.

The common practices to the text /graph separation are briefly depicted in the following. By string extraction and consequently texts and graphs separated is one of the popular methods. The string extraction approaches can be technically categorized as connected component analysis[4~6], and run-length smoothing [7]. One typical approach using component analysis is proposed by [4]. The principle components of the algorithm are the generation of connected components and the application of the Hough transform in order to group components into logical character strings which then are separated from the graphs. Another approach proposed by [5] is also based on connected component analysis. They used a translation and rotation–invariant attribute to cluster connected components. Then they estimated the orientation of a text string by maximum likelihood estimation method (MLE). They could correctly detect the occurrence of overgrouping. But all of their methods handle line graphs only.

S. Imade *et al.*[8] utilizes segmentation and classification method for separating a document image into printed character, handwritten character, photograph, and painted image regions. They first binarized the original image, then divided it into blocks of 8*8 pixels, and every block was reduced to one element. Finally

they used Neural Network to do training by luminance level and gradient vector direction in every block. But all of the above algorithms could not deal with degraded documents.

In [1], it gives an Agent Growing Method (AGM) to handle degraded historical documents. In the case that documents consisting of text and background only, AGM is shown to be very effective and efficient. In this algorithm, partial histogram equalization is first applied to the image to enhance the contrast. Then agent breeding and diffusion techniques are used to identify background pixels and therefore text pixels are located. However, if there are graphs in the document, the algorithm will fail. So we propose an algorithm to identify all the possible graph pixels and then apply AGM to the rest of the document which now has only text- and background-pixels. This paper is organized as follows. The proposed algorithm is described in Section 2. In Section 3, some examples of outputs from our algorithm and some existed algorithms are presented. Section 4 concludes the paper. Fig.1 is the flow of our method.

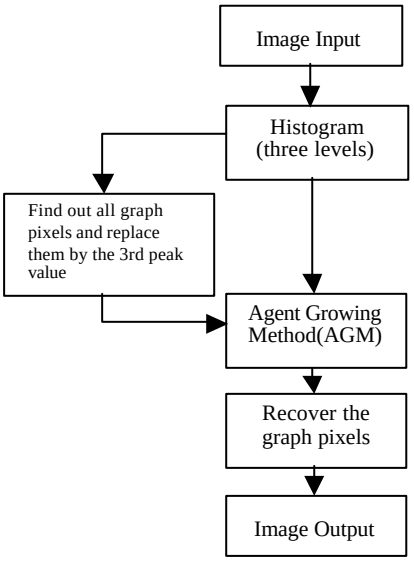


Fig.1 Flow chart of the proposed algorithm

2. THE ALGORITHM

2.1 Histogram analysis and smoothing

A gray-valued document in general consists of texts, graphs if there is any, and background for the rest. Usually, text pixels have small gray values (toward 0) with large standard deviation (SD), and background pixels have large gray values (toward 255) with small SD. Although graph pixels do not have these properties but most of them are connected together.

We first analyze the histogram to help us comprehending the structure of the original image. We want to divide the gray values (0 ~ 255) into 3 levels according to the corresponding histogram such that pixels with gray values in level 1 or 3 are most likely to be text or background respectively. In order to do so we need the information of major peaks and valleys of the histogram. Two simple methods– Laplician Sign method[12] and pyramid data structure[13] – are used three times respectively [20] to do the smoothing and find out the major peaks and valleys. In most of documents it will result in 3 peaks and 2 valleys as in Fig.2(b). Notice that as in Fig.2(b), due to 3 times of pyramid structures, those peaks and valleys will be multiplied by 2^3 to get the corresponding true values for p_1, p_2, p_3, v_1, v_2. Then 3 levels will be [0, v_1], [v_1+1, v_2], [v_2, 255]. In Fig.2(c~e) we show pixels corresponding to 3 levels with their original gray values except in level 3 are shown in reverse.

1) pixels at Level1 (Fig. 2(c)): most pixels will be text pixels, so within a mask (11x11) we use SD (>35) and the numbers of pixels in the mask to determine whether they are text pixels. If there are too many pixels (>half) then they are more likely to be graph pixels ,on the other hand if it has less pixels (≤ 2) then they may be noise.

2) pixels at Level2 (Fig.2(d)): most of them are the residues of graph pixels or text pixels. We use the property of connectedness to distinguish text and graph pixels. Connected component analysis (CCA) is also used to exclude elongated marginal noises and residues of text pixels from graphs pixels.

3) pixels at Level3 (Fig.2(e)): most of them are background pixels, some are the residues of graph pixels. In this level we try to locate the residues of graphs. By comparing the graph pixels recognized from previous 2 levels, if there are many graph pixels in the neighborhood of a pixel in level 3 then it is considered to be a graph pixel too.

In Fig.3-(1), (b) shows the graph pixels found from level 1&2 (shown in black), (c) shows the text found from level 1&2, (d) shows the result of (b) after CCA. In (h), we use morphological opening to preserve the connectedness of the graph.

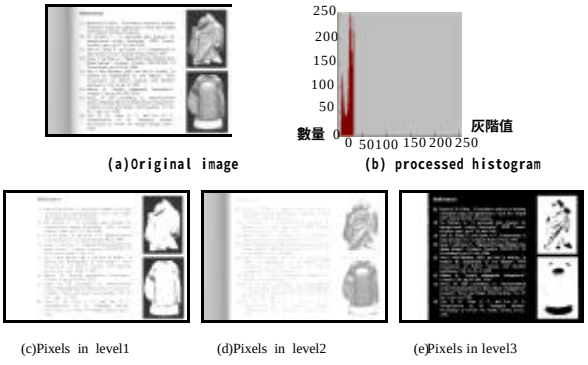


Fig.2 Pixels in three levels

2.2 Agent growing method

Agent growing method (AGM)[1] is an algorithm proposed by Yen, Shih. It can handle degraded historical documents (only Text/background) very well, so we use the algorithm to fine out the text pixels. Hence after graph pixels are all located then the gray values of these levels are replaced by the 3rd peak value (p_3) (see Fig.3-1(f))so to treat them as background pixels, then AGM can be applied to do the rest of the job.

AGM is to identify background pixels of the document, and therefore a success agent means it satisfies a certain criterion and is a background pixel. Two of the most important behaviors of an agent are breeding and diffusion [15][16]. Breeding is to breed more agents from 4-neighbors of a parent agent and diffusion is when an agent fails to satisfy the criterion (so it is a text pixel) then it will jump to the nearby neighborhood to seek for more success agents (if there is any) so the inner holes in letters, like D, O, P, can be found as background. The details of AGM can be found in [1]. After AGM is completed, the original gray values of the graph pixels are recovered , then we get the final result as in Fig.3-1(g).

In some documents, the number of the text pixels is small, or graph pixels may present several typical gray values. Therefore after smoothing process of the histogram, the processed histogram may have only 2 peaks (1 valley) or ≥ 4 peaks (3 valleys). In the former, we let 3 levels to be [0, k],[k+1, v_1], [v_1+1, 255] where k is the gay value that it accumulates 50% of data among data fall within p_1 and v_1. As in the latter we simply let 3 levels to be [0, v_1], [v_1+1, v_k], [v_k+1, 255] if it has (k+1) peaks (k valleys).

3. EXPERIMENTAL RESULTS

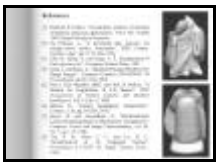
We select several degraded documents from books, papers, and newspapers with degradations caused by various sources. Our method will be compared with global binarization: Otsu's algorithm[16], and local binarization:

Niblack's algorithm[14], Agent-Growing Method. In Fig.3-(1), a partial of a paper copy with luminous marginal noise (2) paper with marginal illumination, (3)dated newspaper with a picture, (4)dated newspaperwithout picture, and (5) the inside cover of a book with word embedded in dark background. The following facts are observed:

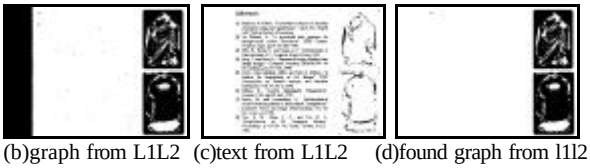
1) The global segmentation method, like Otsu's, can not solve the local variance problem, since it is very possible to have higher gray values for some text pixels than those for other part of the background pixels. Like image(2), pixels will be forced to be text pixels or background pixels, so some marginal noise will be defined to text pixels.
2) Local dynamic threshold, like Niblack's, shows the worst result if the document contains graphs due to the size of the mask (15x15 as recommended in [14]) is smaller than most the graph sizes. But in Fig.3-(4), it has text only, Niblack's performs almost as well as our method.
3) The AGM give a better result than Ostu's and Niblack's Method in most of the cases. However, it is not good enough when the image include graph as in Fig.3-(1).
4) Our proposed method shows the best result of all. It not only can find out text pixels but also preserves the original graphs.

4. CONCLUSION

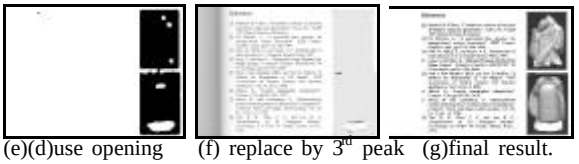
We have presented an simple algorithm for extracting text pixels and graph pixels from degraded documents which background of a severely noise or illumination change and this algorithm is robust to the fonts, language types and any degradation errors in the text. The method uses histogram analysis to distinguish pixels to be background or text in order to find graph pixels. Then replace them by the 3rd peak value in original image. Agent-breeding and agent- diffusion are used to identify background pixels then we get the possible text pixels. Finally, we combine the clear text pixels and the graph pixels with their original gray value . The experimental result of the proposed algorithm has been compared with some binarization algorithms and is shown to have a better result especially for those severely degraded documents.



(a) original



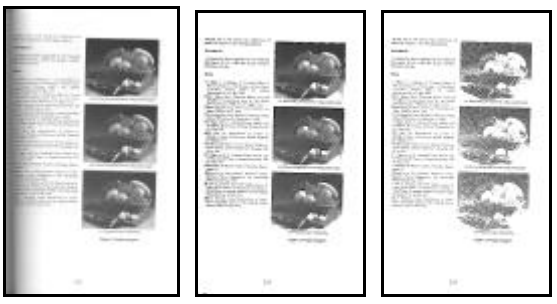
(b)graph from L1L2 (c)text from L1L2 (d)found graph from l1l2



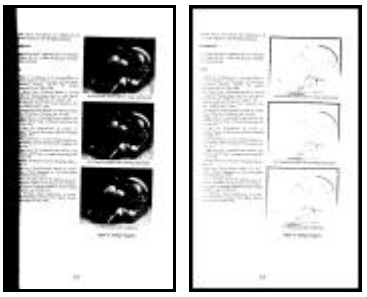
(e)(d)use opening (f) replace by 3rd peak (g)final result.

(h) Using AGM only (i)Ostu's method (h)Niblack's method

Fig.3 -(1)

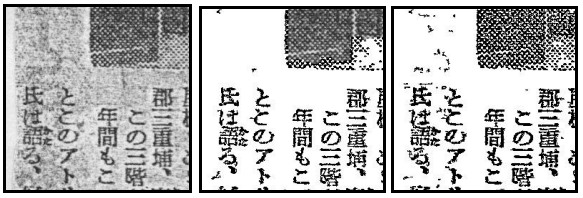


(a) Original (b) Proposed method (c) AGM

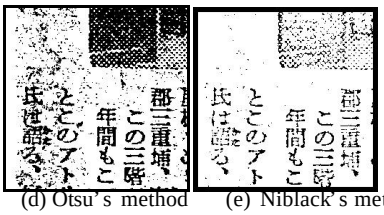


(d) Otsu's method (e) Niblack's method

Fig.3-(2)



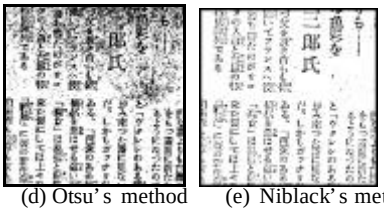
(a) Original (b) Proposed method (c) AGM



(d) Otsu's method (e) Niblack's method

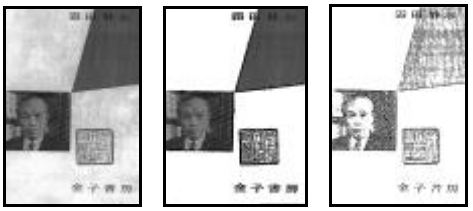
Fig.3-(3)

(a) Original (b) Proposed method (c) AGM

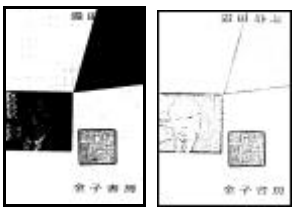


(d) Otsu's method (e) Niblack's method

Fig.3-(4)



(a) Original (b) Proposed method (c) AGM



(d) Otsu's method (e) Niblack's method

Fig.3-(5)

Fig.3 Experimental results

5. REFERENCES

[1] S. H. Yen , M. C. shih , "Histogram Document Reconstruction", SCI 2000 and ISAS 2000, June 2000, pp. 365 – 370.

[2] P.K. Sahoo, S. Soltani, and A. K. C. Wong, "A Survey of Thresholding Techniques", Computer Vision, Graphics, and Image Processing 41, 1988, pp. 233 - 260.

[3]Oivind Due Trier. and Anil K. Jain. ,"Goal –Directed Evaluation of Binarization Methods", IEEE Tran. on Pattern Anal. And Machine Intel., Vol.17, No.12, December 1995, pp. 1191 - 1201.

[4]L. A. Fletcher, R. Kasturi ,"A Robust Algorithm for Text String Seperation from Mixed Text/Graphics Images", IEEE Trans. on Pattern Anal. And Machine Intel., Vol.10, No.6, November 1988,pp.910-918.

[5]S. He , N. Abe, "A Clustering-Based Approach to the Separation of Text Strings from Mixed Text/Graphics documents", IEEE Processing of ICPR 96', 1996, pp. 706 - 710.

[6]C. L. Tan , B. Yuan , W. Huang , Q. Wang , Z. Zhang, "Text/Graphics Separation using Agent-based Pyramid Operator ".

[7]Yibing Yang ,Hong Yan, "An adaptive logical method for binarazation of degraded document images" , Pattern Recognition 33,2000,pp.787-807.

[8]S. Imade , S. Tatsuta, T. Wada "Segmentation and Classification for Mixed Text/Image Documents Using Neural Network", IEEE, 1993, pp. 930 - 934.

[9] J. Suavely , T. Seppanen, S. Happakoski and M. Pietikainen, "Adaptive Document Binarization", Document Analysis and Recognition, Proceedings of Fourth International Conference on Vol. 1, 1997, pp. 147 - 152.

[10] C. L. Tan , B. Yuan , W. Huang , Q. Wang , Z. Zhang, "Text/Graphics Separation using Agent-based Pyramid

Operator ".
[11]Oivind Due Trier. and Anil K. Jain. ,"Goal –Directed Evaluation of Binarization Methods", IEEE Tran. on Pattern Anal. And Machine Intel., Vol.17, No.12, December 1995, pp. 1191 - 1201.
[12] S. Rodtook, Y. Rangsanseri , "Adaptive thresholding of Document Images Based on Laplacian Sign ", IEEE, 2001, pp.501-505.
[13]T. Jiang, M.B .Merickel, E.A. Parrish,JR. ,"Automated Threshold Detection Using a Pyramid Data Structure ", IEEE, 1988, pp. 689 - 692.
[14] J. Suavely , T. Seppanen, S. Happakoski and M. Pietikainen, "Adaptive Document Binarization", Document Analysis and Recognition, Proceedings of Fourth International Conference on Vol. 1, 1997, pp. 147 - 152.
[15] J. Liu , Y. Y. Tang and Y. C. Cao, "An Evolutionary Autonomous Agents Approach to Image Feature Extraction", IEEE Trans. on Evolutionary Computation, Vol. 1, No.2, July 1997, pp. 141 - 158.
[16] N. Otsu, "A Threshold Selection Method from Gray-level Histograms", IEEE Trans. on System, Man, and Cybernetics, Vol. 9, No. 1, January 1979, pp. 377 – 393.

[16] J. Liu and Y. Y. Tang, "Adaptive Image Segmentation With Distributed Behavior-Based Agents", IEEE Trans. on Pattern Anal.
[17] Rafael C. Gonzalez,Richard E. Woods, "Digital Image Processing ", Addison Wesley Inc. Pub.,
[18]Pierre Soille, "Mopholagical Image Analysis" ,Springer Inc. Pub.
[19]Jain, R., R. Kasturi , B. G. Schunck, "Machine Vision", McGraw-Hill,1995
[20]Yi-Jin Chen, "The Extraction of Text/Graphs from Degraded Documents", Master thesis, Department of Comp. Sci. & Inf. Eng., Tamkang Univ. 2002.