

The Extraction of Text/Graphs from Degraded Documents*

Shwu-Huey Yen, Yi-Jin Chen, Hui-Jen Lin and Chia-Jen Wang

Department of Computer Science and Information Engineering, Tamkang University

Tamsui, Taiwan, R.O.C.

e-mail: shyen@cs.tku.edu.tw

Abstract

This paper presents a method for improving the quality of degraded documents by noise removal and text enhancing. Histogram of a degraded document is analyzed to find out the approximate ranges of gray-value for text-, graph-, (i.e. photographs), and background-pixels. After the graph-pixels are identified, they are replaced by the background pixels. Agent-growing method [1] is then applied to smooth the noisy background and a document with clear readable condition for text and background is obtained. At last, graph pixels are recovered to get the final result such that the degraded document now has the text in much better quality and photographs preserved if there is any. Experiments to verify the efficacy of the proposed method and comparison to some existing techniques are also presented.

1. Introduction

Many old dated historical documents, e.g. manuscripts of old literatures, dated newspaper, magazines, are usually very

fragile. As a result, they are kept in a safe and well protected place rather than sharing these historical documents with the general public. After years past, these dated documents often showed a severely degraded condition even if they are well kept. If we can have the original form, clean and clear, of these valuable documents together with their electronic forms (the way they are now) available in the digital library, then they will become a real historical treasure reachable to everyone through the world wide web.

In [1], an agent-growing method (AGM) is proposed to handle degraded historical documents by recovering degraded documents into clean and readable conditions. In the case that a document with line graphs, for the number of pixels in a line graph is far small comparing with the number of pixels in the background, AGM also works well in quality improvement. However, if there are photographs in the document, then algorithm will fail. In this paper, we propose a scheme to successfully identify the photograph pixels and thus extend the AGM for noise removal and text enhancing.

There are a lot of researches about text and graphs separation [2~6]. Texts and graphs separated via string extraction is one of the popular methods. The string extraction approaches can be technically categorized as

* This work is supported by National Science Council, Taiwan, R.O.C., NSC 91-2213-E-032-016.

connected component analysis (CCA) [2~4], run-length smoothing [5], neural network [6], etc. The algorithm proposed in [2] is a typical example of the CCA. The main idea is by producing connected components following by Hough transform to group components into logical character strings, which then separated from the graphs. Another approach proposed by [3] is also based on CCA. They use translation and rotation-invariant attributes to cluster connected components. Then the orientation of a text string is estimated by maximum likelihood estimation method. They can correctly detect the occurrence of over-grouping. But all of these methods handle line graphs only. S. Imade *et al.* [6] utilizes segmentation and classification method for separating a document into printed characters, handwritten characters, photograph, and painted image regions. They first binarize the original image, and divide it into blocks of 8*8 pixels, and every block is reduced to one element. Finally, a Neural Network is trained by luminance level and gradient vector direction in every block. All of the above algorithms could not deal with degraded documents, thus they are not suitable to our problem.

The rest of the paper is organized as follows. Section 2 explains the proposed algorithm. AGM is also briefly introduced. In Section 3, the proposed algorithm has been tested on a set of degraded documents with or without photographs and comparison to some existing techniques are also presented. Finally, concluding remarks are cited in Section 4. In the following discussions, we use graph to represent photograph. Fig.1 is the flow of our method.

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. 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.

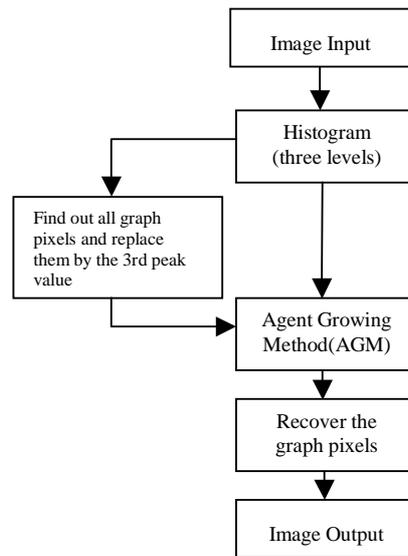


Fig.1 Flow chart of the proposed algorithm

We first analyze the histogram to help us comprehending the structure of the document image. The goal is to divide the gray values of the histogram into 3 levels 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, Laplacian Sign method [8] and pyramid data structure [9], are used three times alternatively for histogram smoothing as well as the major peaks and valleys locating [7]. 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 structure operations, those peaks and valleys will be multiplied by 2^3 to get the corresponding values for p_1, p_2, p_3, v_1, v_2 in the original histogram. Then 3 levels are $[0, v_1], [v_1+1, v_2], [v_2+1, 255]$ and p_1, p_2, p_3 play the roles of typical gray values for text, graph, and background respectively. In Fig.2, a document containing marginal noise as well as photograph, we show pixels corresponding to 3 levels with their original gray values except pixels in *level 3* shown in reverse (c~e). We observe that:

1) Most of pixels at *Level1* are text pixels as in Fig. 2(c). In order to confirm those pixels are indeed text, for a given mask (11x11), it should have a large SD (>35) and the numbers of pixels are not too many and not too small. If there are too many pixels ($>$ -half) then they are more likely to be graph pixels, on the other hand, if it has less than three pixels then they may be noise.

2) Most of pixels at *Level2* are graph pixels and the residues of text pixels as in Fig.2(d). We find out that residues of text pixels form small and noised-like connected components and graph pixels usually connect to each other since they are embedded in a photograph. CCA is applied to distinguish text and graph pixels according to the above discussions. CCA is also used to eliminate elongated marginal noises by examining the dimensions of the components.

3) Most of pixels at *Level3* are background pixels, some are the residues of graph pixels as in Fig.2(e). In this level we try to locate the residues of graphs. For any pixel in *level 3*, if there are many graph pixels recognized from previous 2 levels in the neighborhood then it is considered to be a graph pixel due to connectedness of graph pixels.

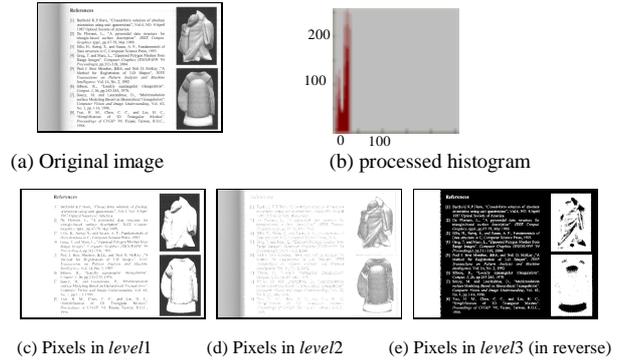


Fig.2 Pixels in three levels

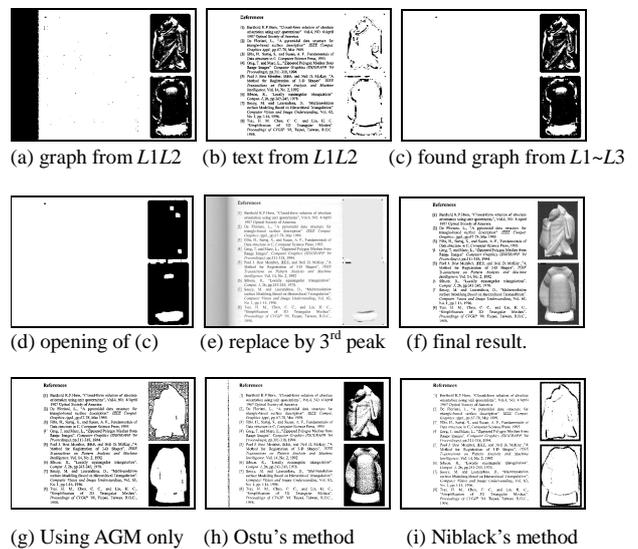


Fig. 3. Steps of proposed algorithm and experimental results.

In Fig.3, (a~d) shown in black, (a) shows the graph pixels found from *level 1 & 2* before CCA, they contain marginal noise and text residues. (b) shows the texts together with residues of texts found from *level 1 & 2* after CCA, since edge pixels inside the photograph also have the property of large SD and not too many or small number of pixels within a mask, they are recognized as text pixels as well. (c) shows the result of (a) after CCA and the residues of the graphs

from *level 3*. We can see that marginal noise and residues of texts are removed after CCA. In order to include edge pixels inside the graph, a morphological opening is applied on (c). As in (d), now most of graph pixels are identified. For those graph-pixels being recognized as background, since they are bright and homogeneous similar to background, they will be replaced by a gray value of 255 later in our algorithm which, as a result, is similar to the way they are originally.

2.2. Agent growing method

AGM is a method, using the fact that the number of text pixels usually is 8% ~ 15% on a document image though 20% is chosen in AGM for safety, first spread out the lower 20% of gray values to be $[0-t]$, $t = \max\{\text{the gray value of 20 percentile of histogram, 150}\}$, in order to exaggerate the difference of the background- and text-pixels. Since background pixels are bright and homogeneous, AGM is to identify background pixels, text pixels are found once all background are located. It will select the first generation “agents” from background pixels according to their own gray values as well as they must be brighter than the average gray value of the corresponding masks (9x9) they are in. The criterions are made to make sure the selected agents must be background. From the first generation agents, called “parents”, they will check whether any of their neighbors (: offspring) are similar to them. If so, this offspring is a success agent and otherwise a failure agent, then offspring becomes parent and continues the process. This is called “breeding”. If an offspring is a failure agent, then a “diffusion” will happen, i.e. its parent will jump to nearby neighborhood of this failure agent to check whether those nearby pixels are similar to the parent, if it is a success then

this nearby neighbor becomes a parent and continues otherwise it stops, “dead” so called for this pixel. The whole procedure will stop once all pixels are checked and every pixel will only be checked once. For those “success” or “parent” agents are exactly background pixels and they will be replaced by a gray value of 255, and those “failure” or “dead” agents are text consequently and they will be replaced by a gray value of 0. Diffusion is to make sure background pixels inside the holes like words: D, O, P, etc, will be located too. The details of AGM can be found in [1].

For any degraded document, following steps as mentioned in 2.1., most of graph pixels can be recognized as in Fig.3(d). Now replace gray values of all those graph pixels by the 3rd peak value (p_3) (see Fig.3(e)) since p_3 is the typical gray value for background. The resulted image turns out to have only text and background pixels. AGM can be applied now to do the job of noise removal and text enhancing. After AGM is completed, the original gray values of the graph pixels are recovered, then we get the final result as in Fig.3(f).

In some documents, the number of the text pixels is very small, or graph pixels may exhibit 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 gray value that it accumulates 50% of data among data fall between p_1 and v_1 . As in the later 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 and discussions

Our method are compared with global binarization:

Otsu's algorithm [11], local binarization: Niblack's algorithm [10], and Agent-Growing Method. Several degraded documents are selected from books, papers, and newspapers with degradations caused by various sources: in Fig.3, a copy of a journal paper with luminous marginal noise; in Fig.4-(1), a journal paper with marginal illumination which has a tilt to the right; Fig.4-(2) is a dated newspaper with a picture; Fig.4-(3) is a dated newspaper without picture; and Fig.4-(4) is the inside cover of a book with words embedded in dark and bright background at the same time. 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 Fig.4-(1), pixels will be forced to be text pixels or background pixels, so some marginal noise will be recognized as 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 [10
- 3)) is smaller than most the graph sizes. But in Fig.4-(3), it has text only, Niblack's performs almost as well as our method.
- 4) AGM gives a better result than Otsu's and Niblack's method in most of cases. However, it is not good enough when the image include graph as in Fig. 3 or Fig.4-(1).
- 5) Our proposed method shows the best result of all. It not only enhances the text pixels but also preserves the original graphs.

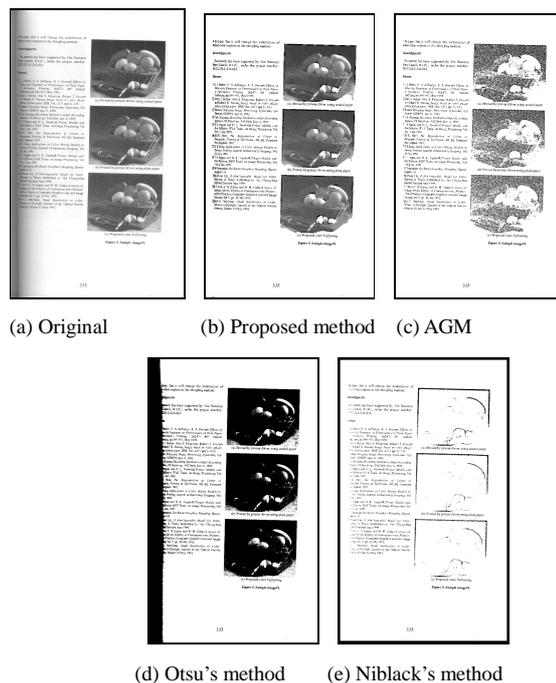


Fig.4-(1)

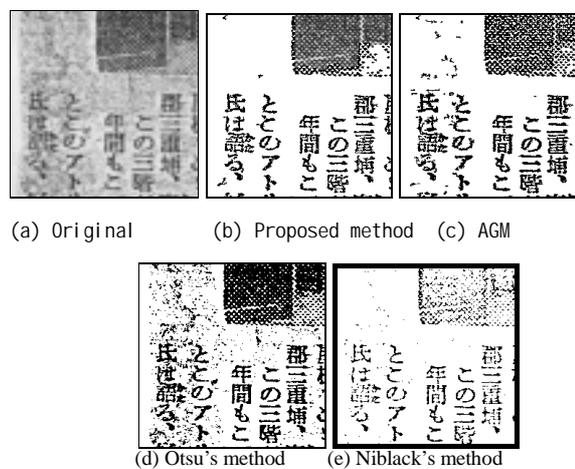
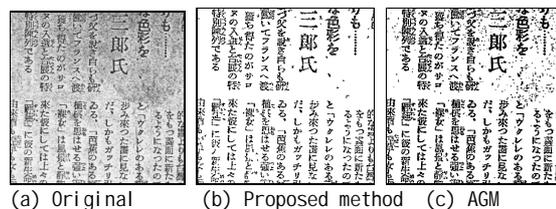


Fig.4-(2)



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

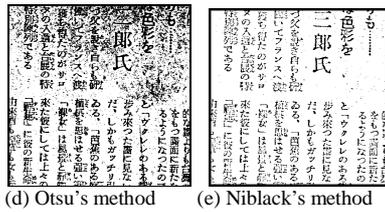


Fig.4-(3)

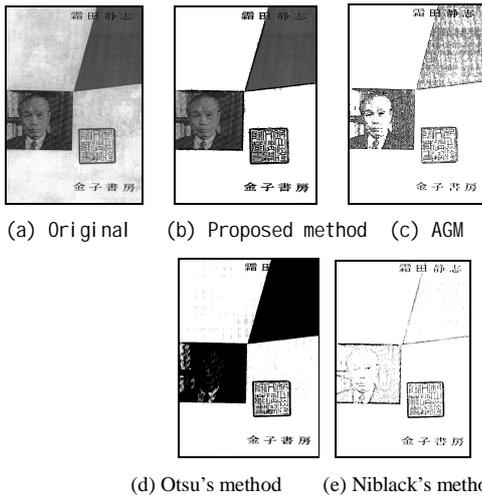


Fig. 4-(4)

Fig.4 Experimental results

4. Conclusion

We have presented an algorithm for identifying graph pixels from degraded documents thus it can improve the quality of the documents by noise removal and text enhancing. This algorithm is robust to the fonts, language types and any degradation errors in the text. The method uses histogram analysis to determine whether pixels are graph pixels. Then replace them by the gray value of the typical background pixels in the original image. AGM is applied on the resulted image now has only text and

background pixels. Finally, we combine the clear text pixels in white background and the graph pixels with their original gray values. The experimental results have been compared with some existing algorithms. Our method is shown to have a better result especially for those severely degraded documents.

5. References

- [1] S. H. Yen, M. C. Shih, "Historical Document Reconstruction", *SCI 2000 and ISAS 2000*, June 2000, pp. 365 – 370.
- [2] L. A. Fletcher, R. Kasturi, "A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images", *IEEE Trans. on Pattern Anal. And Machine Intel.*, Vol.10, No.6, November 1988, pp.910-918.
- [3] 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.
- [4] C. L. Tan, B. Yuan, W. Huang, Q. Wang, Z. Zhang, "Text/Graphics Separation using Agent-based Pyramid Operator". *Document Analysis and Recognition*, 1999. Proceedings of the Fifth International Conference on ICDAR '99. , pp.169 -172
- [5] Yibing Yang, Hong Yan, "An adaptive logical method for binarization of degraded document images" , *Pattern Recognition* 33, 2000, pp.787-807.
- [6] S. Imade, S. Tatsuta, T. Wada "Segmentation and Classification for Mixed Text/Image Documents Using Neural Network", *IEEE*, 1993, pp. 930 - 934.
- [7] Yi-Jin Chen, "The Extraction of Text/Graphs from Degraded Documents", *Master thesis*, Department of Comp. Sci. & Inf. Eng., Tamkang Univ. 2002.
- [8] S. Rodtook, Y. Rangsanseri, "Adaptive thresholding of Document Images Based on Laplacian Sign", *IEEE*, 2001, pp.501-505.
- [9] T. Jiang, M.B.Merickel, E.A. Parrish, JR., "Automated Threshold Detection Using a Pyramid Data Structure", *IEEE*, 1988, pp. 689 - 692.
- [10] J. Suavely, T. Seppanen, S. Happakoski and M. Pietikainen, "Adaptive Document Binarization", *Proceedings of Fourth International Conference on Document Analysis and Recognition*, Vol. 1, 1997, pp. 147 - 152.
- [11] 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.