

Document image preprocessing based on optimal Boolean filters

Win-Long Lee, Kuo-Chin Fan*

Institute of Computer Science and Information Engineering, National Central University, Chung-Li 32054, Taiwan, ROC

Received 22 March 1999; received in revised form 2 August 1999

Abstract

In this paper, optimal Boolean filters are applied to enhance the binary document images corrupted with uniform noise or uniformly distributed distinct graphical patterns in the background. The performance and operation theory of optimal Boolean filters against other competitive techniques are compared. Experimental results show that the Boolean filters outperforms the morphology approach in extracting the text from overlapped text/background images. The feasibility of trained Boolean filters is also confirmed by experimental results in the case where the original image is not available. © 2000 Elsevier Science B.V. All rights reserved.

Zusammenfassung

In diesem Artikel werden optimale Boole'sche Filter angewandt, um Bilder binärer Dokumente aufzubereiten, deren Qualität durch gleichförmiges Rauschen oder durch gleichverteilte ausgeprägte grafische Muster im Hintergrund beeinträchtigt ist. Die Leistungsfähigkeit und die Theorie der Wirkungsweise optimaler Boole'scher Filter wird denjenigen anderer vergleichbarer Techniken gegenübergestellt. Experimentelle Ergebnisse zeigen, daß dies Boole'schen Filter das morphologische Verfahren zur Extraktion von Text aus Bildern übertreffen, wenn Text und Hintergründe überlappen. © 2000 Elsevier Science B.V. All rights reserved.

Résumé

Dans cet article, des filtres booléens optimaux sont appliqués pour améliorer les images de documents binaires corrompus par un bruit uniforme ou par des motifs graphiques distincts uniformément distribués dans l'arrière-plan. La performance et la théorie de ces opérations sont comparées à d'autres techniques compétitives. Des résultats expérimentaux montrent que les filtres booléens dépassent l'approche morphologique pour l'extraction de textes d'images où textes et arrière-plans se recouvrent. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Optimal Boolean filters; Document image enhancement; Text extraction

1. Introduction

The Optical Character Recognition (OCR) system is a convenient entry system for the construction of a digital document database. Yet,

* Corresponding author. Tel.: 886-3-4227151, ext. 4453; fax: 886-3-4222681.

E-mail address: kcfan@ncu.edu.tw (K.-C. Fan)

the enhancement of document images, such as extraction of text characters from background and removal of image noise, is a prerequisite before the OCR system can then be applied. Accordingly, various techniques that can be and are usually applied to enhance document images have been developed. Connected components analysis [12,15] and thresholding [3,5,17] are the traditional techniques to segment the text from backgrounds. But, these techniques do not work if the objects and background pixels cannot be distinguished by their gray level values. A morphology approach [1,7] has also been adopted to remove the noise and extract the texts from overlapped text/background binary document images. But the approach is rather complicated and needs much background knowledge about the morphology theory. Recently, an approach [13] using the complementary similarity measure is proposed to recognize the characters on graphical designs. The algorithm deals with irregular background graphical designs which is beyond our discussion.

In this paper, optimal Boolean filters are adopted to enhance the binary document images and extract the texts from overlapped text/background binary document images. The performance of nonlinear filter in image processing has been extensively studied for decades. However, the design of optimal nonlinear filters has not been well defined until the problem of optimal stack filters initialized by Coyle et al. [16]. They introduced the class of stack filters and found the connection between the stack filters and positive Boolean functions. In [2], Coyle and Lin defined the optimal stack filters under the mean absolute error (MAE) criterion and the problem of designing optimal stack filters was transformed into that of finding optimal positive Boolean functions. Since then many different approaches have been proposed to the finding of optimal stack filters [4,8,10,18]. Most of them belong to the method of training approach since the problem of finding optimal stack filters is considered as an Integer Linear Programming problem.

An extension class of stack filters called Boolean filters was introduced by [6]. The procedures to design optimal Boolean filters have been discussed by Tabus et al. [14]. In this paper, we will concen-

trate on the designing and extending the application of Boolean filters to binary document image processing. The problem of designing optimal Boolean filters is defined under the MAE criterion. In our approach, the statistical measurement between the observed images and desired images is the main computational task. The MAE of a Boolean filter is represented in terms of the error incurred by the input vectors. In this way, the optimal Boolean filter can be obtained immediately.

Usually, we are given a corrupted image to be enhanced but the original image is not available. Hence, it is impossible to find the optimal filter for this corrupted image. However, it is possible to find an optimal filter of a corrupted image from another given set of images assuming that the noise can be regenerated and is irrelevant to the content of images. The given set of images includes an original image and its corrupted version. This is the concept of training filter, being considered as of practical use is the optimal filter theory. The feasibility of trained filter is also investigated throughout the experiments.

The rest of this paper is organized as follows. In Section 2, we will briefly review the definition of Boolean filters. In Section 3, the problem of finding optimal Boolean filters is studied and the relationship of the MAE of Boolean filters and the cost function of the input vectors is derived. The applications of Boolean filters in the area of image processing including binary document image enhancement and text extraction from overlapping text/background images will be presented in Section 4. Real images are tested to evaluate the performance of the optimal Boolean filters. Finally, conclusions are given in Section 5.

2. Boolean filters

The Boolean filters are defined on Boolean function possessing the threshold decomposition property. Let $BF_f(\cdot)$ denote a Boolean filter specified by a Boolean function $f(\cdot)$, X the input gray scale image, T the thresholding function, and $T_k(X)$ the thresholded binary image of X thresholded at gray level k . The threshold decomposition property of

Boolean filters can be expressed as:

$$\text{BF}_f(X) = \sum_{k=1}^{M-1} f(T_k(X))$$

where $(M-1)$ is the largest possible value of x_i , $x_i \in \{0, 1, 2, \dots, M-1\}$. Owing to the threshold decomposition property, the design, analysis and realization of Boolean filters can be reduced to the binary domain.

In our work, the geometrical representation of Boolean functions is adopted. According to the True and False entries of the truth table, the 2^n input vectors of an n -variable Boolean function can be classified into two subclasses, one containing the input vectors corresponding to the True entries of truth table which is called on-set, the other is called off-set which contains the input vectors corresponding to the False entries. The Boolean functions can be completely specified by the on-set [11] and so does the Boolean filters.

3. Optimal Boolean filters under the MAE criterion

In this section, the problem of finding optimal Boolean filters is revisited and the mean absolute error (MAE) is adopted as the error criterion in determining the optimal Boolean filter. The MAE of a Boolean filter can then be represented in terms of the total error incurred by the input vectors of the on-set. In consequence, the optimal Boolean filter can be found immediately. Furthermore, the relationship between the Boolean filters, stack filters and WOS (weighted order statistic) filters can be revealed based on the structure of the on-set. The optimal design complexities of these filters are also analyzed in this section.

For a Boolean filter BF_f where f is the Boolean function defining the Boolean algebra of the Boolean filters, the mean absolute error (MAE) of the Boolean filters is defined as the mean absolute error between the desired image Z and the output of Boolean filters with the observed image X serving as the input

$$\text{MAE}(\text{BF}_f) = E[|Z - \text{BF}_f(X)|] \quad (3.1)$$

where $E[\cdot]$ is the expectation operator.

The optimization problem can be stated as the finding of a Boolean filter which minimizes Eq. (3.1). According to the threshold decomposition property, the MAE of a Boolean filter can be reduced to the sum of the decision errors made by the Boolean filters on each level of thresholded binary images. That is,

$$\begin{aligned} \text{MAE}(\text{BF}_f) &= E[|Z - \text{BF}_f(X)|] \\ &= E\left[\left|\sum_{k=1}^M T_k(Z) - \sum_{k=1}^M f(T_k(X))\right|\right] \\ &= \sum_{k=1}^M E[|T_k(Z) - f(T_k(X))|] \end{aligned}$$

The MAE can be further represented as the decision errors incurred by the input vectors. Now, let us define a cost function, $\text{cost}(b)$, as the decision error incurred by $f(b)$ for deciding a 1 when seeing input vector b .

$$\text{cost}(b) = C(0, b) - C(1, b).$$

In implementation, the cost coefficients $C(0, b)$ and $C(1, b)$ are computed by statistical measurement. That is, the $C(0, b)$ is the number of occurrences of the input vector b appearing in the observed image when the desired output of this vector is 0, $C(1, b)$ is the number of occurrences of the input vector b appearing in the observed image when the desired output of this vector is 1. As a result, the value of $\text{cost}(b)$ is a statistical measurement which is equal to the difference in the number of occurrences between the desired outputs of the vector b . Since b is the observed input vector in the observed thresholded binary images, 0 and 1 are the corresponding pixel values in the desired thresholded binary images. If $\text{cost}(b) > 0$, then it means that the number of occurrences of the desired 0 outputs of the vector b is larger than that of the desired 1 outputs of the vector b . In such a case, it is helpful to include the input vector b in the off-set so as to obtain a filtered output image that is expected to be the desired image. Similarly, if $\text{cost}(b) < 0$, then the b should be included in the on-set in order to obtain a desired filtered output image.

By the definition of MAE, the MAE of a Boolean filter can be reformulated as

$$\text{MAE}(\text{BF}_f)$$

$$= E \left\{ \sum_{k=1}^{M-1} \left[\sum_{b \in \text{on}(f)} C(0, b) + \sum_{b \in \text{off}(f)} C(1, b) \right] \right\}.$$

Since by definition $C(0, b) = \text{cost}(b) + C(1, b)$, it will yield

$$\text{MAE}(\text{BF}_f)$$

$$= E \left\{ \sum_{k=1}^{M-1} \left[\sum_{b \in \text{on}(f)} \text{cost}(b) + \sum_{b \in \text{on}(f)} C(1, b) + \sum_{b \in \text{off}(f)} C(1, b) \right] \right\}$$

$$\text{As } \sum_{b \in \text{on}(f)} C(1, b) + \sum_{b \in \text{off}(f)} C(1, b)$$

$$= \sum_{\forall b} C(1, b), \text{ it yields}$$

$$\text{MAE}(\text{BF}_f)$$

$$= E \left\{ \sum_{k=1}^{M-1} \left[\sum_{\forall b} C(1, b) + \sum_{b \in \text{on}(f)} \text{cost}(b) \right] \right\}. \quad (3.2)$$

The value of $\sum_{\forall b} C(1, b)$ is a constant since it is actually equal to the sum of all pixel values of the desired image. The MAE of Boolean filters is therefore decided by $\sum_{b \in \text{on}(f)} \text{cost}(b)$. The value of $\sum_{b \in \text{on}(f)} \text{cost}(b)$ is not a constant but depends on the total cost of input vectors in the on-set. Different on-set defines different Boolean filter and results in different values of $\sum_{b \in \text{on}(f)} \text{cost}(b)$. Consequently, the optimal on-set which defines the optimal Boolean filter is the on-set such that the total cost of input vectors is minimal among all on-sets. Hence, the optimal Boolean filter can be obtained immediately after the computation of the cost function based on Eq. (3.2).

4. Experiment results

In this experiment, the optimal Boolean filters are applied to the applications of document enhancement and text extraction from overlapping text/background images. The window size of the optimal Boolean filters is 3×3 . In finding the optimal Boolean filters, the cost parameters of the 512

input vectors are computed from the original image and its corrupted version. All input vectors with negative cost can then specify the optimal Boolean filter. In this approach, the original image is a prerequisite in finding the optimal Boolean filter. However, it is impractical in real applications. Hence, a question arises: What to do if the original image is not available?

The concept of trained stack filter in [9] is adopted in this experiment. The trained stack filter is an image enhancement method that obtains an optimal stack filter from an image and its corrupted version and then applies this trained stack filter to another corrupted image to suppress the noise.

4.1. Document image enhancement

Suppose that we get a noisy binary document image corrupted by additive noise during transmission. In that case, the noise cannot be eliminated by thresholding because the gray level of noise is the same as those of the text characters. Experimental results show that the optimal Boolean filter is efficient in eliminating the additive noise of document images. Let us compare the performance of the optimal Boolean filter with that of the median filter and a heuristic algorithm which enhances document images by deleting the isolated points. Fig. 1(a) is the original 256×256 document image. Fig. 1(b) is the testing noisy document image corrupted by adding 10% additive noise, Fig. 1(c) is the enhanced image generated by a 3×3 median filter, Fig. 1(d) is the enhanced image generated by deleting isolated pixels. Here, the isolated pixel is defined as a pixel that does not connect to any other pixels. Fig. 1(e) is the enhanced image generated by the 3×3 optimal Boolean filter. The results show that the optimal Boolean filter gives the best performance.

Besides, it is interesting to note that the median filter shows very good performance in eliminating the additive noise of gray scaled images. However, it does not work well for the additive noise embedded in binary document images. In this experiment, the median filter eliminates not only the noise but also the text characters as shown in Fig. 1(c). The reason is that the median filters belong to the class of Boolean filters and only work well in relatively restricted cases. As to the optimal Boolean filter, it

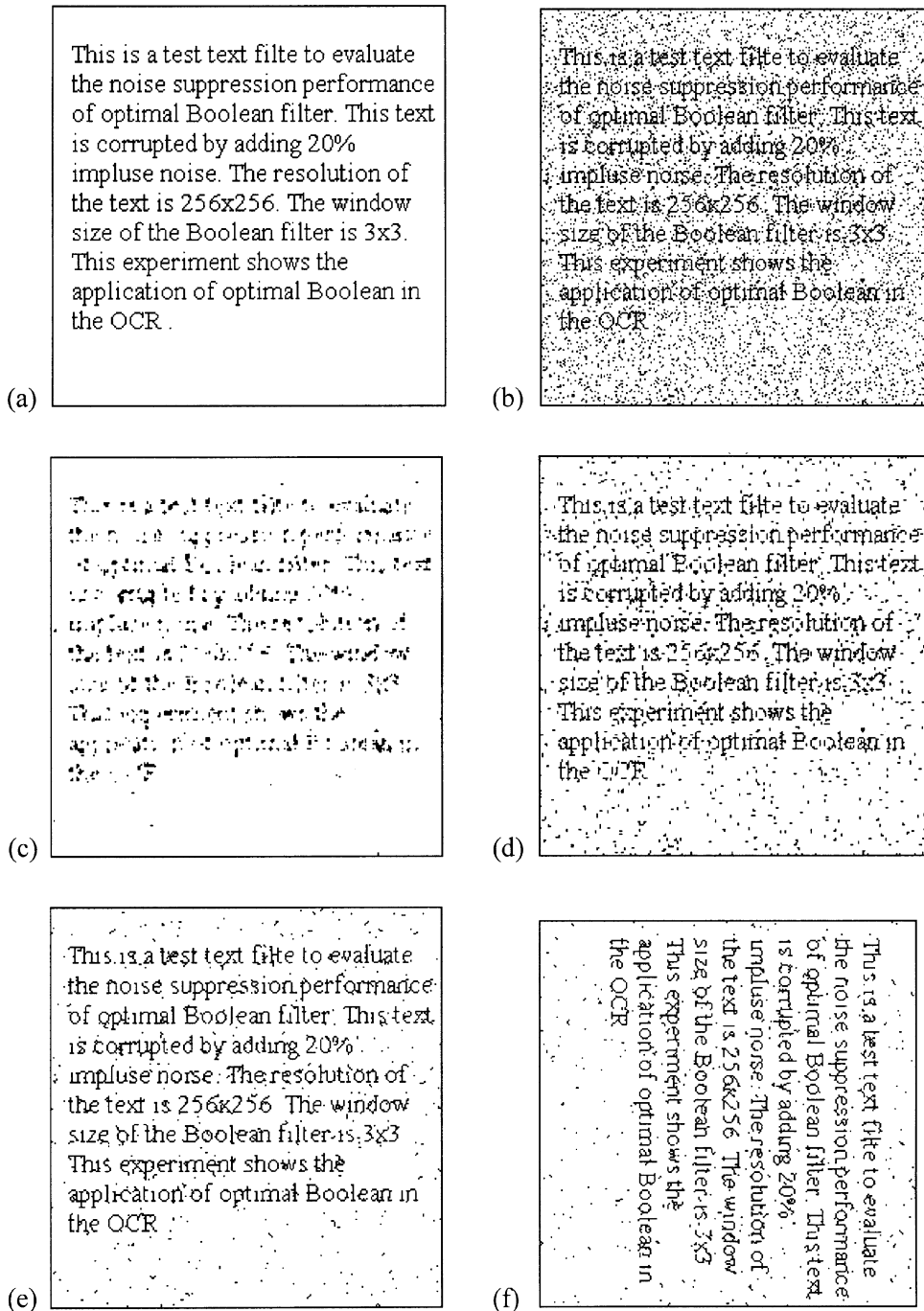


Fig. 1. Comparison of the performance of document image enhancement between different approaches. (a) The original document image; (b) the testing noisy document image corrupted by adding 10% additive noise; (c) the enhanced image generated by a 3×3 median filter; (d) the resulting image generated by deleting isolated pixels; (e) the enhanced image generated by the 3×3 optimal Boolean filter; (f) the enhanced image generated by applying the 3×3 optimal Boolean filter on the noisy image rotated by 90° .

is the global optimal one and can always have a better performance than that of median filters. In this experiment, the cardinality of the optimal on-set is equal to 312. In other words, the optimal Boolean filter passes the 312 input vectors that are considered as the patterns of characters and removes another 200 input vectors that are supposed to be the noise patterns. There are totally 512 input vectors for the window size of 3×3 . On the contrary, the on-sets of median filters are fixed. A median filter of window size 3×3 will always pass the fixed 256 input vectors and remove another fixed 256 input vectors.

The Boolean filtering is translation invariant because the filtering process is essentially a pattern matching process. The Boolean filtering operated on the characters are almost rotation invariant since many of the characters are symmetric. Fig. 1(f) is the enhanced image generated by applying the 3×3 optimal Boolean filter on the noisy image rotated by 90° , which has only little variance as compared to Fig. 1(e). Besides, the output images generated by the optimal Boolean filter as the noisy image is rotated by 180° and 270° are also very close to those of Fig. 1(e).

The trained Boolean filter is also effective in this case. We apply the optimal Boolean filter trained by Fig. 1(a) and Fig. 1(b) to the noisy document image Fig. 2(a). Fig. 2(b) is the result generated by

the trained optimal Boolean filter which demonstrates the capability of the filter in noise removal while preserving the text characters.

4.2. Text extraction from overlapping text/background image

Occasionally, texts will be printed over uniformly distributed graphical background to beautify the layout of articles or attract the attention of readers. Though it is not difficult for humans to read these characters, the extraction of the characters from overlapping background images is still a prerequisite to facilitate the process of Optical Character Recognition (OCR). A morphological approach [7] that based on mathematical morphology has been successively used to solve this kind of problem. We found that the optimal Boolean filter is also effective in separating the graphic background and text characters. A testing document image with the background composed of uniform dots as shown in Fig. 3(a). Fig. 3(b) shows that the optimal Boolean filter can extract the text characters completely. Moreover, Fig. 3(c) is a testing document image that contains overlapping text, texture and 10% random noise. The optimal Boolean filter can still extract the text as shown in Fig. 3(d). Fig. 4 is another example generated by the optimal Boolean filter in the extraction of text characters while the

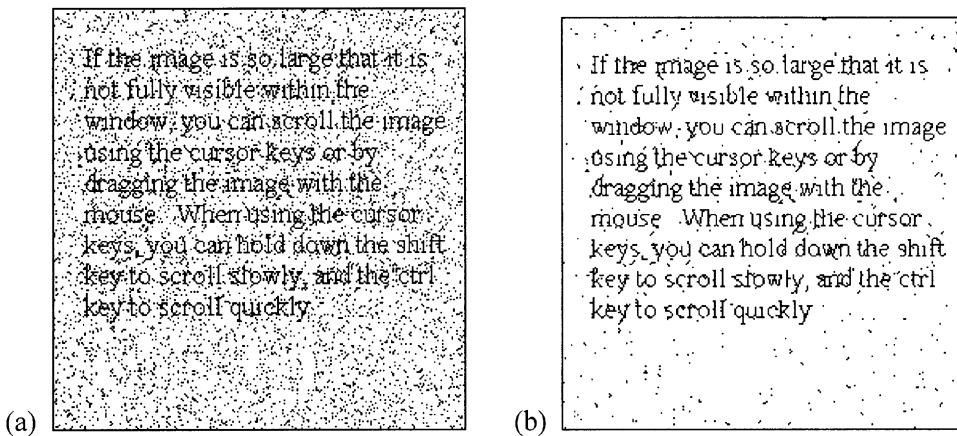


Fig. 2. The experimental result of the trained Boolean filter. (a) The testing noisy document image; (b) the enhanced image generated by the 3×3 trained Boolean filter trained from Fig. 1(a).

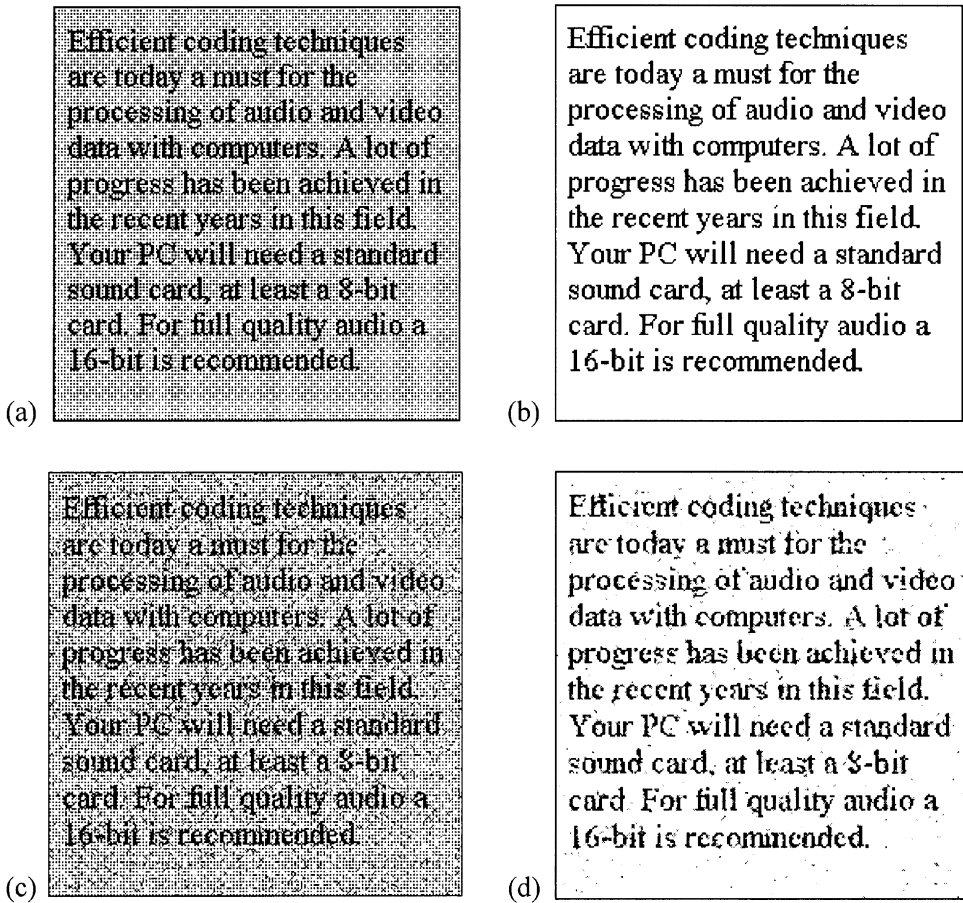


Fig. 3. (a) A testing document image with overlapping text/background; (b) the output image generated by the optimal Boolean filter; (c) a testing document image with overlapping text, texture and 10% random noise; (d) the output image generated by the optimal Boolean filter.

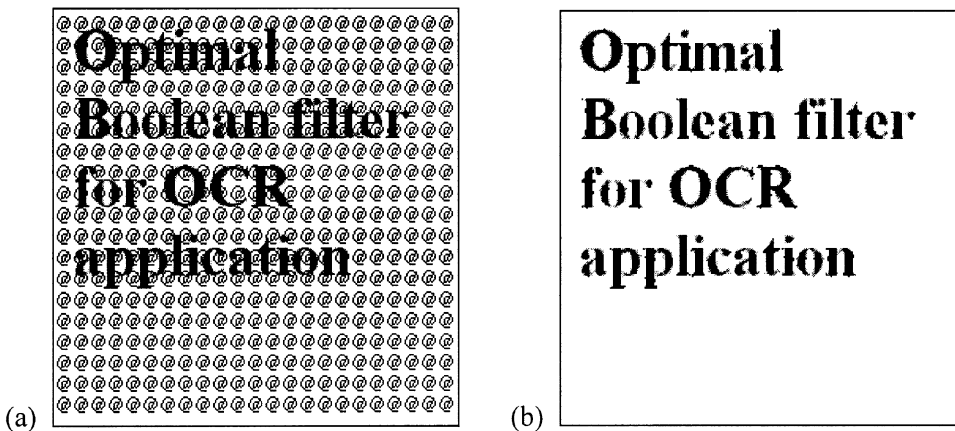


Fig. 4. The result generated by the optimal Boolean filter in the extraction of text characters from overlapping text/background image. (a) A testing document image with overlapping text/background; (b) the output image generated by the optimal Boolean filter.

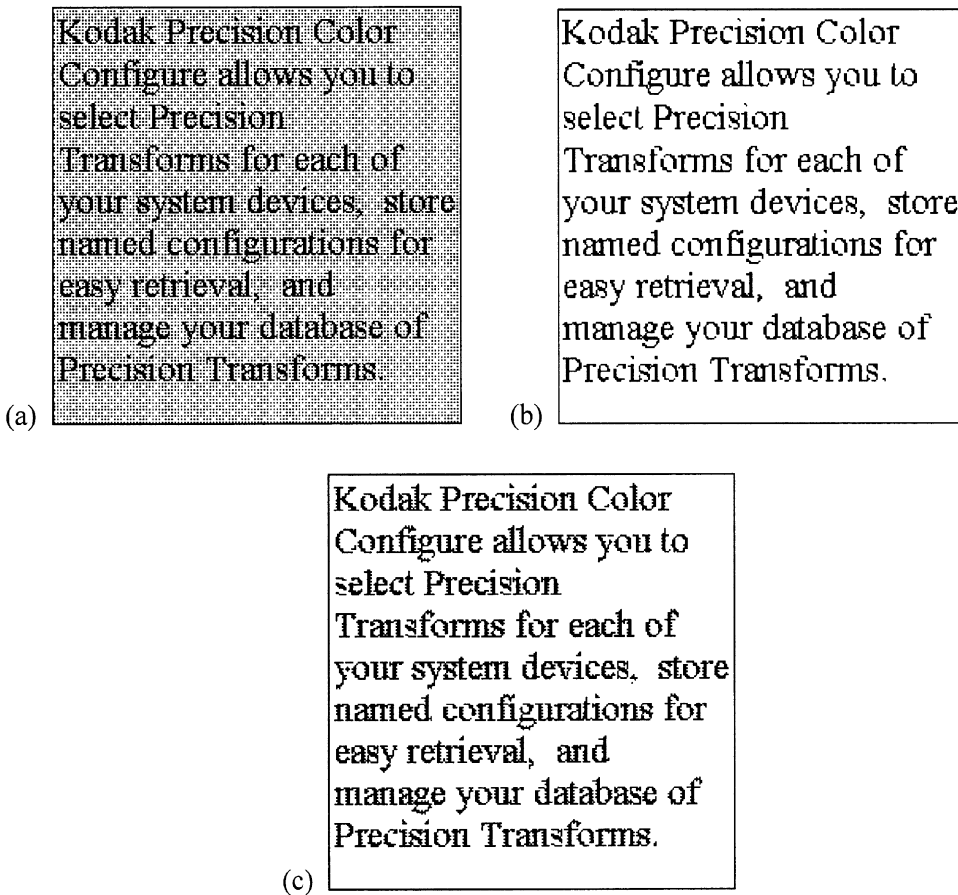


Fig. 5. (a) A testing document image with overlapping text/background; (b) the output image generated by the trained Boolean filter trained from Fig. 3; (c) the result of text extraction generated by the morphological approach.

background is composed of graphic patterns. Besides, the trained Boolean filter is also feasible in performing this task. The optimal Boolean filter is also efficient in extracting the text when the background is made up of periodic patterns as is shown in Fig. 5. Fig. 5(b) is the output image generated by the trained Boolean filter of Fig. 3. Fig. 6(b) is the output image generated by the trained Boolean filter of Fig. 4.

The advantage of the morphology approach compared to the approach of trained Boolean filter morphology, is that it does not require specific training for each type of background pattern/noise. The advantages of the optimal Boolean filter approach are:

1. The design method of optimal Boolean filters is simpler and the extraction process is faster.
2. The extraction of text characters outperforms the approach of morphology. Take Fig. 5(a) as an example, the result generated by the morphological approach contains more noise on the characters as shown in Fig. 5(c). Another example as shown in Fig. 6(c) is the result of text extraction by using the morphological approach to extract the overlapped image of Fig. 6(a).

The common property of the two approaches is that they have to detect the pixel distribution of the background graphic symbol. However, the detecting algorithm is different. The optimal

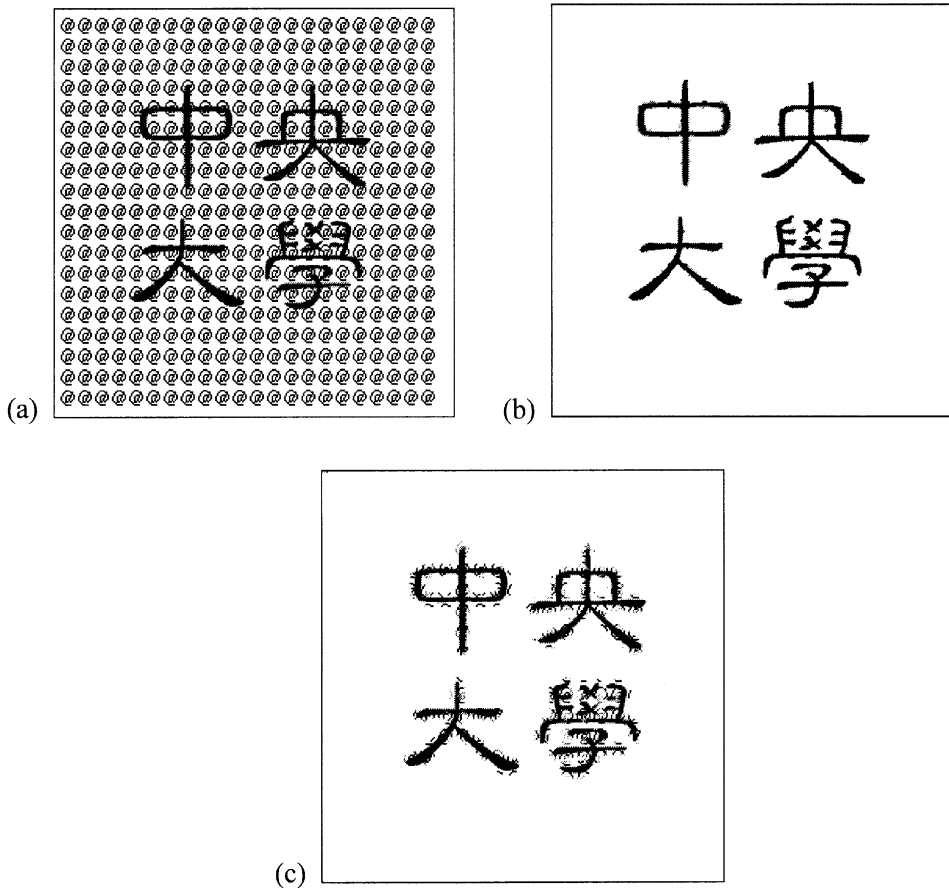


Fig. 6. (a) A testing document image with overlapping text/background; (b) the output image generated by the trained Boolean filter trained from Fig. 4; (c) the result of text extraction generated by the morphological approach.

Boolean filter detects the background distribution by learning from an overlapping text/background image and the same text image without background symbols. The optimal Boolean filter is represented by an on-set which is like a look-up table deciding what kind of input vectors should be “1” and what kind of input vectors should be “0”. Therefore, the text extraction performance of optimal Boolean filters should not be affected by the defects in the periodicity of background symbols. The results as shown in Fig. 7(b) verify this opinion. On the contrary, the morphological approach detects the background distribution by computing the distribution frequency of the background graphic

symbols. The regularity and periodicity of distributed background symbols are prerequisites for the morphological approach. The text extraction performance of the morphological approach as shown in Fig. 7(c) will be degraded by the defects in the periodicity of background symbols. Hence, the morphological filter is not an adaptive filter. Yet, the background extraction procedure of the referred morphological approach [7] is not a morphological filtering process. It requires that the background patterns should be complete and uniform. The background patterns that are uncompleted and non-uniform will not be extracted accordingly.

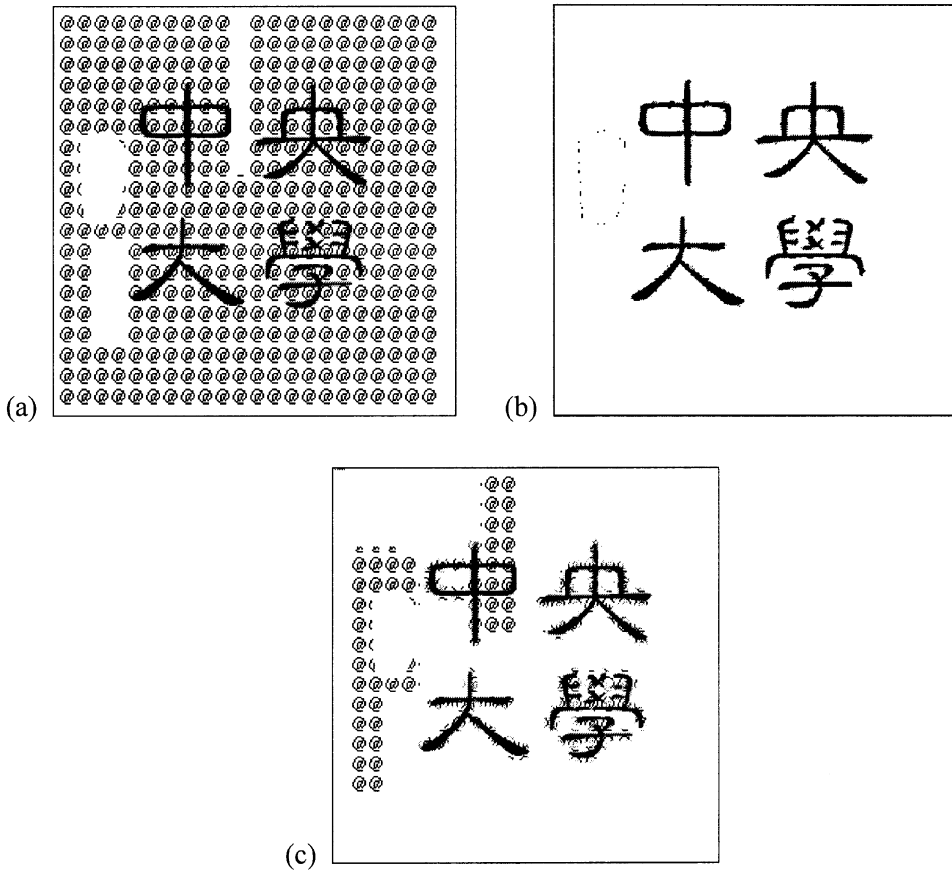


Fig. 7. (a) A testing document image with defects in the background symbols; (b) the output image generated by the optimal Boolean filter approach that is not affected by the defects in the periodicity of background symbols; (c) the text extraction performance of the morphological approach which is degraded by the defects in the periodicity of background symbols.

5. Conclusions

In this paper, the design and applications of optimal Boolean filters are discussed. The good performance of Boolean filters demonstrated in this paper confirms the potential of Boolean filters in binary document image processing. One observation from experimental results reveals the fact that the optimal Boolean filters are effective and robust in suppressing the additive signal no matter whether the noise gets suppressed by the additive signal or the background symbol gets removed by the additive signal. Especially, the experimental results of text character extraction from overlapping

text/background image reminds us that Boolean filters may be suitable for the image segmentation and texture analysis, which becomes our further work in the study of optimal Boolean filters.

References

- [1] M. B. H. Ali, Background noise detection and cleaning in document images, *Proceedings of IEEE International Conference on Pattern recognition*, 1996, pp. 758–762.
- [2] E. J. Coyle, J.-H. Lin, Stack filters and the mean absolute error criterion, *IEEE Trans. Acoust. Speech Signal Process.* 36 (August 1988) 1244–1254.

- [3] H.-S. Don, A noise attribute thresholding method for document image binarization, *Proceedings of the Third International Conference on Document Analysis and Recognition*, 1995, pp. 231–234.
- [4] M. Gabbouj, E. J. Coyle, On the LP which finds a MMAE stack filter, *IEEE Trans. Signal Process.* 39(11) (November 1991) 2419–2424.
- [5] R.C. Gonzalez, R.E. Woods, *Digital Image Processing*, Addison-Wesley, Reading MA, 1992.
- [6] K. D. Lee, Y. H. Lee, Threshold Boolean filter, *IEEE Trans. Signal Process.* 42 (8) (August 1994) 2022–2036.
- [7] S. Liang, M. Ahmadi, M. Shridhar, A morphological approach to text string extraction from regular periodic overlapping text/background images, *CVGIP: Graphical Models Image Process.* 56 (5) (1994) 402–413.
- [8] J.-H. Lin, E. J. Coyle, Minimum mean absolute error estimation over the class of generalized stack filters, *IEEE Trans. Acoust. Speech Signal Process.* 38(4) (April 1990) 663–678.
- [9] J.-H. Lin, Y.-T. Kim, Fast algorithms for training stack filters, *IEEE Trans. Signal Process.* 42(4) (April 1994) 772–781.
- [10] J.-H. Lin, T. M. Sellke, E. J. Coyle, Adaptive stack filtering under the mean absolute error criterion, *IEEE Trans. Acoust. Speech Signal Process.* 38 (April 1990) 938–954.
- [11] S. Muroga, *Threshold Logic and Its Applications*, Wiley, New York, 1971.
- [12] A. Rosenfeld, On connectivity properties of gray pictures, *Pattern Recognition* (16) (1983) 47–50.
- [13] M. Sawaki, N. Hagita, Text-line extraction and character recognition of document headlines with graphical designs using complementary similarity measure, *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (10) (October 1998) 1103–1109.
- [14] I. Tabus, D. Petrescu, M. Gabbouj, A training framework for stack and Boolean filtering – fast optimal design procedures and robustness case study, *IEEE Trans. Image Process.* 5 (6) (June 1996) 809–825.
- [15] Y. Wang, P. Bhattacharya, Image analysis and segmentation using gray connected components, *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, 1996, pp. 444–449.
- [16] P. D. Wendt, E. J. Coyle, N. C. Gallagher, Stack filters, *IEEE Trans. Acoust. Speech Signal Process.* ASSP-34 (4) (August 1986), 898–911.
- [17] J. M. White, G. D. Rohrer, Image thresholding for optical character recognition and other applications requiring character image extraction, *IBM J. Res. Dev.* 27 (4) (1983) 400–410.
- [18] B. Zeng, M. Gabbouj, Y. Neuvo, A unified design method for rank orderstack, and generalized stack filters based on classical Bayes decision, *IEEE Trans. Circuits Systems*, 38 (1991) 1003–1020.