Restoration of Archival Documents Using a Wavelet Technique

Chew Lim Tan, *Member*, *IEEE*,
Ruini Cao, and Peiyi Shen

Abstract—This paper addresses a problem of restoring handwritten archival documents by recovering their contents from the interfering handwriting on the reverse side caused by the seeping of ink. We present a novel method that works by first matching both sides of a document such that the interfering strokes are mapped with the corresponding strokes originating from the reverse side. This facilitates the identification of the foreground and interfering strokes. A wavelet reconstruction process then iteratively enhances the foreground strokes and smears the interfering strokes so as to strengthen the discriminating capability of an improved Canny edge detector against the interfering strokes. The method has been shown to restore the documents effectively with average precision and recall rates for foreground text extraction at 84 percent and 96 percent, respectively.

Index Terms—Document image analysis, wavelet enhancement, wavelet smearing, Canny edge detector, text extraction, image segmentation, bleed-through, show-through, noise cancellation, denoising.

1 Introduction

AN important task in document image analysis is text segmentation [1], [2]. However, this paper introduces a rather different problem of text segmentation, that is, how to extract clear text strings from interfering images originating from the reverse side. The motivation of this research arises from a request from the National Archives of Singapore to restore the appearance of their handwritten archival documents that have been kept over long periods of time during which the seeping of ink has resulted in double images as shown in Fig. 1. Given this problem, we have to separate three classes of objects: the foreground text, interfering strokes from the reverse side, and the background. Usually, the foreground writing appears darker than the interfering strokes. However, there are cases where the foreground and interfering strokes have similar intensities, or worse still, the interfering strokes are more prominent than the foreground.

Many segmentation and binarization approaches have been reported in the literature [2], [3], [4], [5], [6], [7]. These methods aim to extract clear text from either noisy or textured background. However, they deal with one-sided documents and most methods basically assume separable gray scale and/or distinctive features between the foreground and background.

Similar work can be seen in solving the "show-through" problem in scanning duplex printed documents. Don's work segments the double-sided images based on the isolated gray-scale range of interfering images and the noise characteristics [8]. Sharma [9], [10] develops a unique scanning model and an adaptive linear-filtering scheme for removal of show-through using both sides of the document. Our problem, however, is different from show-through in that the interfering strokes are the result of "bleed-through" due to anisotropic absorption and spreading in the paper. As such, corresponding images on both sides may not be completely matched like in the "show-through" situation.

- *C.L. Tan is with the School of Computing, National University of Singapore, 3 Science Drive 2, Singapore, 117543. E-mail: tancl@comp.nus.edu.sg.*
- *R. Cao is with Hotcard Technology Pte Ltd., 2 Jurong East Street 21, #05-30 IMM Building., Singapore 609601. E-mail: caorn@hotcardtech.com.*
- *P. Shen is with the Communication Solution Group, Agilent Technologies, Yishun Ave. 7, No. 1, Singapore, 768923. E-mail: pei-yi_shen@aglient.com.*

Manuscript received 27 Dec. 2000; revised 17 Aug. 2001; accepted 10 May 2002. Recommended for acceptance by L. Vincent.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 113366.

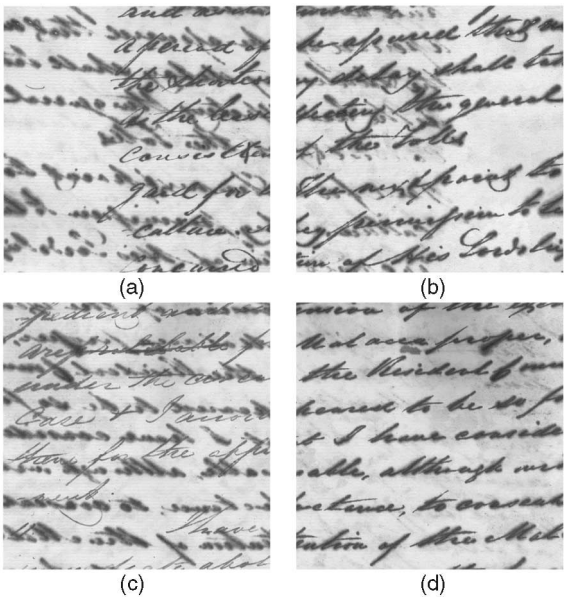


Fig. 1. Sample images: (a) sample1, front side, (b) sample1, reverse side, (c) sample2, front side, and (d) sample2, reverse side.

Related work is seen in the denoising techniques presented by Donoho [11], [12] based on thresholding and shrinking empirical wavelet coefficients for recovering and/or denoising signals. Berkner et al. [13] propose a wavelet-based approach to sharpening and smoothing of images for use in deblurring or denoising of images. Our problem again differs from these works here in that the interfering strokes are not really noise but rather distinctive images in their own right that we want to distinguish and remove.

Finally, Lu et al. [14], Lu [15] presents a similar wavelet method by decreasing the edge contrast and smearing the direct components of the edges with its neighboring pixels. Though his edge-based wavelet image preprocessing method can handle the change of feature coefficients (local maxima) [16], [17], [18], it is still inadequate for our present problem: 1) Due to the "bleed-through" problem discussed earlier, we find different edge positions between the interfering and original strokes on either side. 2) As a result, any mismatch between the interfering strokes observed on the front and their original strokes on the reverse side will result in a mistaken identity of interfering strokes as foreground edges.

In view of the above, we propose a new method which differs from others in the following aspects. 1) We develop an improved Canny edge detector to suppress unwanted interfering strokes [19], [20]. 2) We process the image by using wavelet enhancing and smearing operations to work on the foreground and interfering strokes, respectively. 3) We adopt a set of wavelet enhancement and smearing coefficients in different scales instead of the traditional local maxima reconstruction method.

2 Proposed Method

2.1 Image Matching and Overlay

It is natural that weak foreground strokes may not necessarily seep into the reverse side (Fig. 1d). On the other hand, interfering strokes must have been originated from strong foreground strokes on the reverse side. Thus, we match both images from either side of a page by hand. Let $f(x, y)$ denote the k bits per pixel gray-scale front images, and $b(x, y)$ the reverse side image of the same page, where x and y represent the row and the column, respectively. The two images have the same dimension $M \times N$. An overlay operation is carried out as follows:

Fig. 2. Overlay results of images in Fig. 1: (a) sample1, front side, (b) sample1, reverse side, (c) sample2, front side, and (d) sample2, reverse side.

1. Invert the reverse side image:

$$invert(b(x,y)) = 2^k - 1 - b(x,y). \quad (1)$$

2. Flip the inverted reverse side image and superimpose it on the front image such that corresponding strokes on either side are matched to effect an image subtraction:

$$a(x,y) = flip(invert(b(x,y))) + f(x,y), \quad (2)$$

where $flip()$ means flipping the image horizontally resulting in its mirror image:

$$flip(b(x,y)) = b(x, N-y). \quad (3)$$

3. Scale the resultant image:

$$c(x,y) = curve\left(\frac{a(x,y) - \min(a(x,y))}{\max(a(x,y)) - \min(a(x,y))} * (2^k - 1)\right), \quad (4)$$

where $curve()$ is a nonlinear transform that shrinks low intensity values to make the subtracted strokes in the overlay operation less prominent.

$$y = curve(x) = 2^k - 1 - \sqrt{(2^k-1)^2 - x * x}. \quad (5)$$

Fig. 2 shows the results of overlay and nonlinear curve processing.

2.2 Enhancement and Smearing Features

Comparing Fig. 1 and Fig. 2, we could see that the overlay preprocessing has weakened most of the interfering strokes. Though some of the foreground strokes have also been affected, the major portion of the foreground strokes remains intact. These foreground strokes, though somewhat impaired, now serve as seeds to start the following enhancement and smearing processes. The idea now is to detect the foreground strokes on the front and enhance them using wavelets. The detected and binarized strokes from the foreground overlay image form the "enhancement feature image." Similarly, we detect the foreground strokes on the reverse side to locate their corresponding interfering strokes on the front so as to smear these interfering strokes using wavelets. The detected and binarized strokes from the reverse side overlay image result in the "smearing feature image." Figs. 3a and 3b show the detected enhancement/smearing features of sample 1 and sample 2, respectively. Iterative enhancement and smearing processes are then carried out on the original front image using the enhancement and smearing feature images to identify candidate strokes. The same process could be done on to the reverse side where the reverse side will be treated as the front.

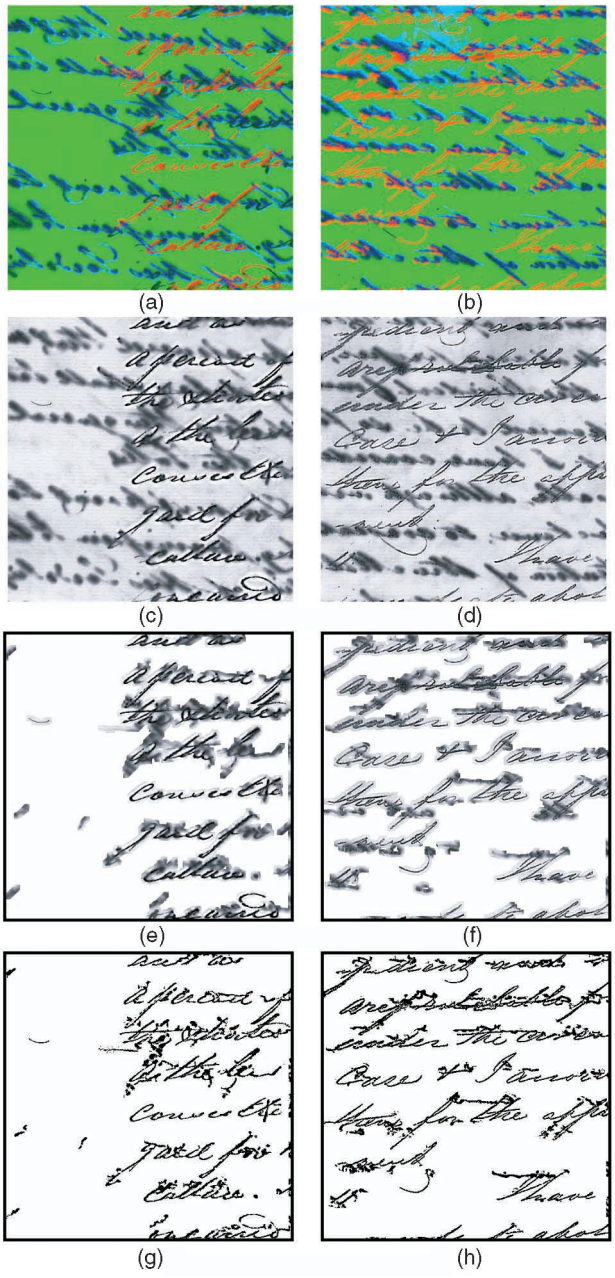


Fig. 3. (a) Enhancement/smearing features of Fig.1a, (b) enhancement/smearing features of Fig.1c, (Note: Multichannel in (a) and (b): green, original background; red: enhancement features; blue: smearing features); (c) enhanced/smeared image of Fig. 1a; (d) enhanced/smeared image of Fig. 1c, (e) segmentation result of (c), (f) segmentation result of (d), (g) final binarized image of (e), and (h) final binarized image of (f).

Let $E(x,y)$ be the enhancement feature image and $S(x,y)$ be the smearing feature image. Both have the same dimension $M \times N$ as the front image $f(x,y)$ and the reverse side image $b(x,y)$. The enhancement and smearing features may be described as follows:

$$E(x,y) = \begin{cases} 255, detected & stroke\ on\ f(x,y) \\ 0, & otherwise, \end{cases}$$

$$S(x,y) = \begin{cases} 255, detected & stroke\ on\ b(x,y) \\ 0, & otherwise. \end{cases} \quad (6)$$

To detect the foreground strokes on either side so as to determine $E(x,y)$ and $S(x,y)$, an improved Canny edge detection algorithm with an orientation filter and orientation constraint [19], [20] is first adopted to pick up the edges of the foreground strokes based on the observation that the interfering strokes are not as sharp as the foreground strokes. The orientation filter favors foreground strokes that are predominantly slanting at a particular angle against interfering strokes slanting at a different angle. The orientation constraint helps minimize erroneous edges caused by nearby interfering strokes. The resultant edges then serve as loci to recover streaks of gray-level images of a predetermined width from the foreground overlay images. The recovered streaks are binarized using an adaptive Niblack's threshold [21] to give $E(x,y)$ and $S(x,y)$, respectively.

2.3 Iterative Wavelet Reconstruction

The wavelet decomposition of $f(x,y)$ is written as follows [22],

$$\begin{cases} C_j f(m,n) = (< f(x,y), \Phi_{j,m,n}(x,y) >)_{(m,n)\in Z^2} \\ D_j^1 f(m,n) = (< f(x,y), \Psi_{j,m,n}^1(x,y) >)_{(m,n)\in Z^2} \\ D_j^2 f(m,n) = (< f(x,y), \Psi_{j,m,n}^2(x,y) >)_{(m,n)\in Z^2} \\ D_j^3 f(m,n) = (< f(x,y), \Psi_{j,m,n}^3(x,y) >)_{(m,n)\in Z^2}, \end{cases} \quad (7)$$

where $C_j f(m,n)$ is the wavelet approximation coefficient and $D_j^k f(m,n)$ $(k=1,2,3)$ are the wavelet detail coefficients at scale j of the wavelet decomposition. With the image wavelet representation $Wf(x,y)$, the enhancement feature $E(x,y)$ and smearing feature $S(x,y)$, the iterative wavelet reconstruction may be described as follows:

1. The multiscale decomposition of the foreground. Unlike the traditional wavelet decomposition, the resultant image in each scale and iteration retains the same size as the original foreground image.

$$Wf(x,y) = \{C_J(x,y), D_j^k(x,y), j=0,\ldots,J, k=1,2,3\}. \quad (8)$$

2. Basically, wavelet details D_j^k are related to the high frequency components in the edges which are enhanced and smeared by the enhancement coefficient e_j^k and smearing coefficient s_j^k, respectively, according to $E(x,y)$ and $S(x,y)$ as below:

$$\begin{aligned} Do\{ \quad & if\ E(x,y) == 255\ D_j^k(x,y) = e_j^k D_j^k(x,y); \\ & if\ S(x,y) == 255\ D_j^k(x,y) = s_j^k D_j^k(x,y); \\ \}& for(j=0,\ldots,J; k=1,2,3; x=1,\ldots,M; y=1,\ldots,N), \end{aligned} \quad (9)$$

where e_j^k and s_j^k $\{e_j^k > 1, 1 > s_j^k > 0, j=0,\ldots,J, k=1,2,3,\}$ are set empirically. The enhanced/smeared image $f'(x,y)$ is reconstructed from the modified coefficients.

$$f'(x,y) = inverse\ wavelet\ transform(\{C_J(x,y), D_j^k(x,y), j=0,\ldots,J, k=1,2,3\}). \quad (10)$$

3. To the reconstructed image $f'(x,y)$, apply the wavelet transform again. Note that, in obtaining the inverse of the wavelet transform, the revised D_j^k obtained in (9) are used again.

$$Wf'(x,y) = \{C'_J(x,y), D'^k_j(x,y), j=0,\ldots,J, k=1,2,3\} \quad (11)$$

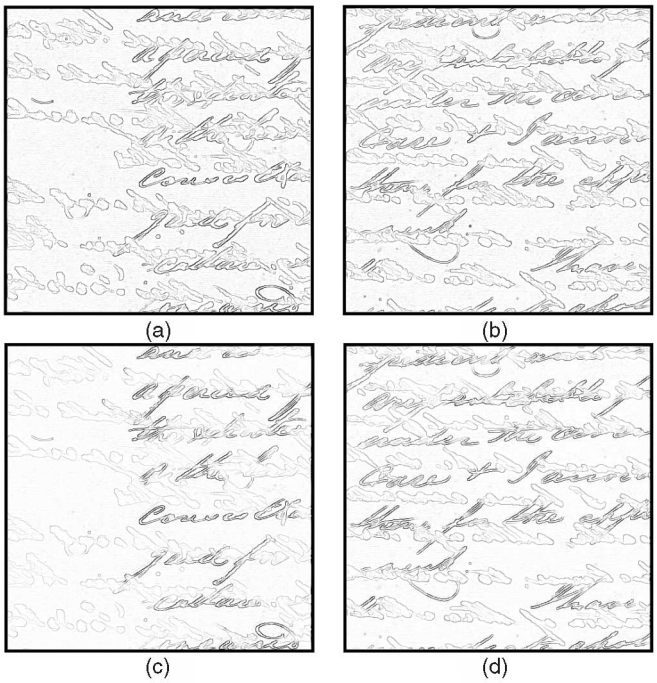


Fig. 4. (a) and (b) Magnitude of gradient of original images shown in Figs. 1a and 1c, respectively. (c) and (d) Magnitude of gradient of enhanced/smeared images shown in Figs. 3c and 3d, respectively.

$$f'(x,y) = inverse\ wavelet\ transform(\{C'_J(x,y), D_j^k(x,y), j=0,\ldots,J, k=1,2,3\}). \quad (12)$$

4. Iteratively, process the wavelet decomposition and reconstruction using (11) and (12), we could get the final enhanced/smeared gray-scale image. The iteration progressively strengthens or weakens the strokes in the reconstructed image through the gradual changes in the edge [16].
5. Clip the final enhanced/smeared image using the following function:

$$f''(x,y) = \begin{cases} 0, & f'(x,y) \le 0 \\ 255, & f'(x,y) \ge 255 \\ f'(x,y), & otherwise. \end{cases} \quad (13)$$

In our implementation, we use Daubechies wavelets due to the wide variety of the singularities of the interfering strokes. We empirically set the filter length at 6. A 3-scale wavelet decomposition procedure is adopted after some experimentation. During the wavelet decomposition, all the subimages keep the same size as the original image. In the reconstruction process, we set 15 to be the maximum number of iterations. Figs. 3c and 3d show the results after 15 iterations. After the final enhancement and smearing, the resultant image is then processed by the same improved Canny edge detector described in Section 2.2 to obtain the final output in Figs. 3e and 3f followed by binarization leading to Figs. 3g and 3h, respectively.

2.4 Robust Threshold Decision in Canny Edge Detector

Figs. 4a, 4b, 4c, and 4d show the edge strength images of the original front images and their enhanced/smeared images, respectively. The gradient's magnitude is converted into the gray-level value. The darker the edge is, the larger is the gradient magnitude. It is obvious from Figs. 4a and 4b that without the enhancement/smearing processes, the edge strengths of strong interfering strokes are similar to that of the foreground strokes. Thus, it is difficult to set a universally valid pair of double thresholds for Canny edge detector. This is so in view of the great variety of the relative strengths between the foreground and interfering strokes among these

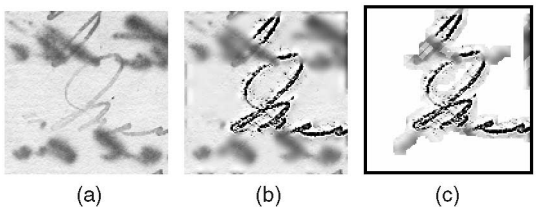


Fig. 5. (a) Weak foreground strokes in a small size window. (b) Enhanced weak foreground strokes and smeared strong interfering strokes. (c) Final extraction results with the weak foreground detected.

archival documents. In fact, it is sometimes even impossible to set one single set of thresholds for the same page due to the variation of strokes intensity across the page. On the other hand, from Figs. 4c and 4d, it is seen that the enhancement/smearing processes have significantly highlighted the foreground strokes against the interfering strokes. Three interesting observations came to light through our experimentation with the enhancement and smearing processes.

First, extremely faint foreground strokes may not be detected as enhancement features and, subsequently, not strengthened in the enhancement process. However, this can be remedied by applying the Canny detector on small subwindows to be later reassembled together (i.e., choosing a smaller value of $M \times N$ mentioned in Section 2.1). Within a small subwindow, the feeble foreground strokes will become comparatively prominent and will be picked up by the relative threshold (in percentage) of the Canny detector. Fig. 5 shows a case in point involving a 128×128 subimage where extremely faint foreground strokes are detected in the enhancement feature image and, subsequently, enhanced in the final result.

Second, the enhancement and smearing processes work with each other to the advantage of our desirable results. Generally, a lower value for the Canny detector's upper threshold is adopted to detect as many features as possible. The enhancement feature image may have erroneously picked up interfering strokes as enhancement features, resulting in a noisy enhancement feature image. However, with the same lower threshold value, more smearing features will also be included in the smearing feature image. Some of the smearing features will be in the overlapped areas (partially or fully)

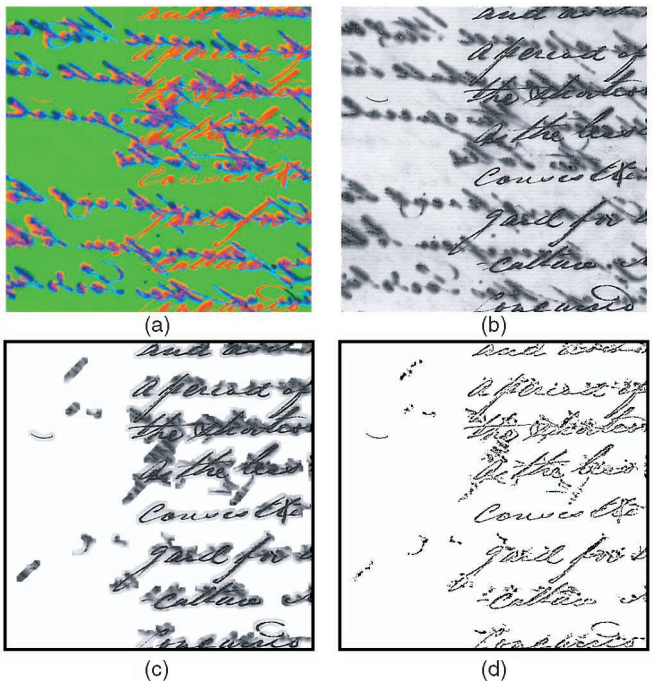
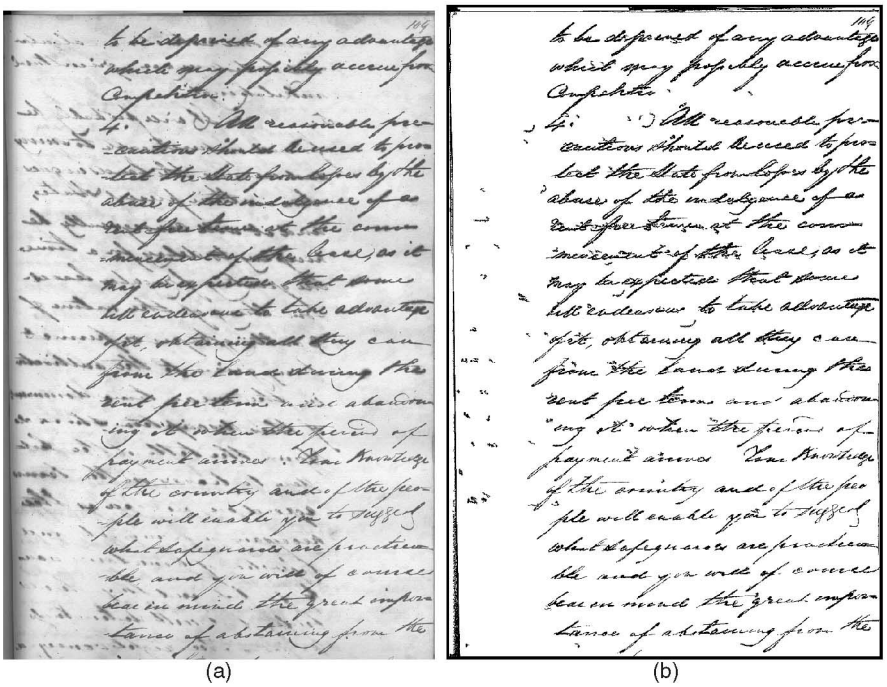


Fig. 6. The robustness of picking up edges. (a) Enhancement/smearing features, red: enhancement features, blue: smearing features, green: original background. (b) Wavelet enhancement/smearing reconstruction. (c) Segmentation results. (d) Final binarized results.

Fig. 7. Sample images in Table 1 and their final segmentation results.



(a) (b)

Fig. 8. (a) Original gray-scale image and (b) its segmentation result (size: $1,146 \times 1,734$).

TABLE 1
Evaluation of the Proposed Method

Image number	1	2	3	4	5	6	7	8	9	10	11	12	Average
Total no. of Words	132	124	103	125	125	123	121	128	112	113	114	114	
Precision	91%	86%	76%	94%	92%	80%	79%	75%	84%	82%	91%	78%	84%
Recall	98%	100%	90%	94%	98%	94%	89%	97%	97%	98%	96%	96%	96%

with the falsely identified strokes in the enhancement feature image. As a result, these false alarms will be eventually suppressed by the subsequent smearing process. The beauty of this property is that as long as a smearing feature covers any part of a mistaken enhancement feature, this false positive will eventually be "smeared" away. The unwanted strokes will finally be sifted out by the cancellation effect of the smearing process. It is also possible that the smearing image contains misclassified edge features which are later mistakenly smeared out. However, these misclassified smearing features should correspond to strong strokes on the front side and will accordingly be detected in the enhancement image. So, the smearing effect will be "cancelled" by the enhancement process. This collaborative nature between enhancement and smearing in fact leads to the third observation, that is, unlike the conventional edge detection, the final detection of the foreground strokes from the enhanced/smeared image is robust to the threshold setting.

Fig. 6a shows many mistaken enhancement features that are always associated with the smearing features. Thus, in this noisy image, the final enhanced/smeared results still have the desired appearance, as shown in Figs. 6b, 6c, and 6d. Comparing Fig. 3g with Fig. 6d, we could see clearly the faint character "C" in the bottom line is recovered in Fig. 6c, while it was lost in Fig. 3e. And the final binarized image Fig. 6d does not induce any extra noise as compared to Fig. 3g.

3 Experiment Results

We tested our system with scanned images of historical handwritten documents given by the National Archives of Singapore. The images were scanned at 150 dpi and saved as TIF format without compression. Most of the images are moderately noisy and were satisfactorily cleaned up. To assess the performance of our method especially for difficult cases, 12 severely interfering images were selected for evaluation. The selected images were visually inspected to assess the readability of the extracted words. Fig. 7 shows all the 12 sample images in cut off strips and the final binary segmentation, while Fig. 8 gives a full view of one of the images. We use *precision* and *recall* defined below [23].

Precision = No. of Correctly Detected Words / No. of All Words detected by the system,

Recall = No. of Correctly Detected Words / Total No. of Words present in the document,

where the total number of words in the document includes all words in the foreground, while the total number of words detected means the sum of the correctly detected words and incorrectly detected words (interfering words). The number of correctly detected words was counted manually. If some characters in a foreground word are lost or not recovered properly, the whole word is considered lost. If parts of characters coming from the back are picked up by the system, the total number of incorrectly detected words will be increased by 1. Precision reflects the performance of removing the interfering strokes and recall reflects the performance of restoring the foreground words. The results in Table 1 show average precision and recall rates of 84 percent and 96 percent, respectively.

4 Conclusion

This paper describes a method for the removal of interfering strokes in archival document images. The wavelet-based enhancement/smearing algorithm performs well even for faint foreground strokes with strong interference. The whole system can restore the appearance of the original documents effectively. A problem encountered presently is in getting a perfect manual over a

between the front and reverse side images due to differences between both images caused by factors like document skews, different scales during image capture, and warped surfaces at books' spine areas. We are currently working on the development of a computer-assisted method to do the image overlay.

ACKNOWLEDGMENTS

This research was supported in part by the Agency for Science, Technology & Research and the Ministry of Education, Singapore, under research grant R252-000-071-303/112. The authors would like to thank the National Archives of Singapore for the use of their archival documents.

REFERENCES

[1] G. Nagy, "Twenty Years of Document Image Analysis in PAMI," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 1, pp. 38-62, Jan. 2000.
[2] R.G. Casey and E. Lecolinet, "A Survey of Methods and Strategies in Character Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 7, pp. 690-706, July 1996.
[3] H. Negishi, J. Kato, H. Hase, and T. Watanabe, "Character Extraction from Noisy Background for an Automatic Reference System," *Proc. Fifth Int'l Conf. Document Analysis and Recognition*, pp.143-146, Sept. 1999.
[4] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. System, Man, and Cybernetics,* vol. 9, no. 1, pp. 62-66, 1979.
[5] Y. Liu, S.N. Srihari, "Document Image Binarization Based on Texture Features," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 5, pp. 540-544, May 1997.
[6] S. Liang and M. Ahmadi, "A Morphological Approach to Text String Extraction from Regular Periodic Overlapping Text/Background Images," *Graphical Models and Image Processing,* vol. 56, no. 5, pp. 402-413, Sept. 1994.
[7] J.M. White and G.D. Rohrer, "Image Thresholding for Optical Character Recognition and Other Applications Requiring Character Image Extraction," *IBM J. Research Development,* vol. 27, no. 4, pp. 400-410, 1983.
[8] H.-S. Don, "A Noise Attribute Thresholding Method for Document Image Binarization," *Proc. Third Int'l Conf. Document Analysis and Recognition,* pp. 231-234, Aug. 1995.
[9] G. Sharma, "Cancellation of Show-through in Duplex Scanning," *Proc. Int'l Conf. Image Processing,* vol. 3, pp. 609-612, Sept. 2000.
[10] G. Sharma, "Show-through Cancellation in Scans of Duplex Printed Documents," *IEEE Trans. Image Processing,* vol. 10, no. 5, pp. 736-754, May 2001.
[11] D.L. Donoho, "Threshold Selection for Wavelet Shrinkage of Noisy Data," *Proc. 16th Ann. Int'l Conf. IEEE Eng. in Medicine and Biology Soc.,* vol. 1, pp. A24-A25, Nov. 1994.
[12] D.L. Donoho, "De-Noising by Soft-Thresholding," *IEEE Trans. Information Theory,* vol. 41, no. 3, pp. 613-627, May 1995.
[13] K. Berkner, M.J. Gormish, E.L. Schwarts, and M. Boliek, "A New Wavelet-Based Approach to Sharpening and Smoothing of Images in Besov Spaces with Applications to Deblurring," *Proc. Int'l Conf. Image Processing,* vol. 3, pp. 797-800, Sept. 2000.
[14] J. Lu, D.M. Healy, and J.B. Weaver, "Contrast Enhancement of Medical Images Using Multiscale Edge Representation," *Optical Eng.,* vol. 33, no. 7, pp. 2151-2161, July 1994.
[15] J. Lu, "Image Deblocking via Multiscale Edge Processing," *Proc. SPIE Wavelet Application in Signal and Image Processing IV,* M.A. Unser, A. Aldroubi, and A.F. Laine, eds., vol. 2825, part 2, pp. 742-751, Aug. 1996.
[16] S. Mallat and S. Zhong, "Characterization of Signals from Multiscale Edges," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 14, no. 7, pp. 710-732, July 1992.
[17] W.L. Hwang, F. Chang, "Character Extraction from Documents Using Wavelet Maxima," *Proc. SPIE: Wavelet Applications in Signal and Image Processing IV,* vol. 2825, part 2, M.A. Unser, A. Aldroubi, and A.F. Laine, chairs/eds., pp.1003-1015, Aug. 1996.
[18] K. Etemad, D. Doerman, and R. Chellappa, "Multiscale Segmentation of Unstructured Document Pages Using Soft Decision Integration," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 1, pp. 92-96, Jan. 1997.
[19] R. Cao, C.L. Tan, Q. Wang, and P. Shen, "Segmentation and Analysis of Double-Sided Handwritten Archival Documents," *Proc. Fourth IAPR Int'l Workshop Document Analysis Systems,* pp. 147-158, Dec. 2000.
[20] C.L. Tan, R. Cao, P. Shen, J. Chee, and J. Chang, "Removal of Interfering Strokes in Double-Sided Document Images," *Proc. Fifth IEEE Workshop Applications of Computer Vision,* pp. 16-21, Dec. 2000.
[21] W. Niblack, *An Introduction to Digital Image Processing,* pp. 115-116, Englewood Cliffs, N.J.: Prentice Hall, 1986.
[22] L. Feng, Y.Y. Tang, and L.H. Yang, "A Wavelet Approach to Extracting Contours of Document Images," *Proc. Fifth Int'l Conf. Document Analysis and Recognition,* pp. 71-74, Sept. 1999.
[23] M. Junker, R. Hoch, and A. Dengel, "On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy," *Proc. Fifth Int'l Conf. Document Analysis and Recognition*, pp. 713-716, Sept. 1999.

Estimating the Intrinsic Dimension of Data with a Fractal-Based Method

Francesco Camastra and Alessandro Vinciarelli

Abstract—In this paper, the problem of estimating the intrinsic dimension of a data set is investigated. A fractal-based approach using the Grassberger-Procaccia algorithm is proposed. Since the Grassberger-Procaccia algorithm performs badly on sets of high dimensionality, an empirical procedure that improves the original algorithm has been developed. The procedure has been tested on data sets of known dimensionality and on time series of Santa Fe competition.

Index Terms—Bayesian information criterion, correlation integral, Grassberger-Procaccia's algorithm, intrinsic dimension, nonlinear principal component analysis, box-counting dimension, fractal dimension, Kolmogorov capacity.

1 INTRODUCTION

PATTERN recognition problems involve data represented as vectors of dimension d. The data is then embedded in the space $\mathbb{R}^d$, but this does not necessarily mean that its intrinsic dimension (ID) is d. The ID of a data set is the minimum number of free variables needed to represent the data without information loss. In more general terms, following Fukunaga [9], a data set $\Omega \subset \mathbb{R}^d$ is said to have an ID equal to M if its elements lie entirely within an M-dimensional subspace of $\mathbb{R}^d$ (where $M < d$).

Estimation of the ID is important for many reasons. The use of more dimensions than strictly necessary leads to several problems. The first one is the space needed to store the data. As the amount of available information increases, the compression for storage purposes becomes even more important. The speed of algorithms using the data depends on the dimension of the vectors, so a reduction of the dimension can result in reduced computational time. Moreover, in the statistical learning theory approach [32], the capacity and the generalization capability of the classifiers depend on ID and the use of vectors with smaller dimension often leads to improved classification performance. Finally, when using an autoassociative neural network [18] to perform a nonlinear feature extraction (e.g., nonlinear principal component analysis), the ID can suggest a reasonable value for the number of hidden neurons.

This paper presents an approach to ID estimation based on fractal techniques. Fractal techniques have been successfully applied to estimate the attractor dimension of underlying dynamic systems generating time series [17]. The literature presents results in the study of chaotic systems (e.g., Hénon map, Rössler oscillator) [22], in the analysis of ecological time series (e.g. Canadian lynx population) [15], in biomedical signal analysis [31], in radar clutter identification [12], and in the prediction of financial time series [24]. Nevertheless, in pattern recognition, fractal methods are mainly used to measure the fractal dimension of an image [13]. As far as we know, the application of fractal approaches to the problem of ID estimation has never been proposed before. The proposed ID estimation method is tested on both artificial and real data showing good results.

The paper is organized as follows: Section 2 presents several techniques for estimating ID. In Section 3, fractal methods are reviewed. The procedure to estimate ID is described in Section 4. In Section 5, some experimental results are reported and, in Section 6, some conclusions are drawn.

- *F. Camastra is with the INFM-DISI, University of Genova, Via Dodecaneso 35, 16146 Genova, Italy. E-mail: camastra@disi.unige.it.*
- *A. Vinciarelli is with the Institut Dalle Molle d'Intelligence Artificielle Perceptive, Rue du Simplon 4, 1920 Martigny, Switzerland. E-mail: vincia@idiap.ch.*

Manuscript received 18 July 2001; revised 14 Jan. 2002; accepted 2 Apr. 2002. Recommended for acceptance by P. Meer.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 114549.