Machine Print Filter for Handwriting Analysis

Siyuan Chen, Sargur Srihari

Machine Print Filter for Handwriting Analysis

Siyuan Chen and Sargur Srihari
Center of Excellence for Document Analysis and Recognition (CEDAR)
University at Buffalo, State University of New York, Buffalo, USA
{schen7, srihari}@cedar.buffalo.edu

Abstract

Existing document decomposition models fail in well separating the zone areas of printed text and handwritten text when they are close or even touching each other. This paper presents a simple and robust algorithm to filter out the printed content in a mixed document. Following the traditional bottom-up approach, the printed text candidates are extracted and detected in the connected component level. Then the relative spatial relation and window-based filter help providing useful information to decide the components to be removed finally. Results with manually cropped signature blocks that contain extraneous printed text show that over 85% of printed text components are removed while preserving handwritten content.

1 Introduction

A business document consists of several physical and semantic layers corresponding to: letter head, printed text, handwritten signature and notes, tables and graphs, etc. The procedure of transforming a paper document to an appropriate electronic representation involves the phases of image enhancement, skew correction, image segmentation, geometric layer analysis and semantic layer analysis [1, 2]. When it comes a clean and well printed document, the document decomposition models developed so far can accomplish the task quite successfully [1, 3, 4]. However in the case of extremely noisy document, segmentation error due to the quality degradation remains an unsolved problem. Basically the methods in document decomposition use discrete connected component as the minimum element over the analysis procedure. Very often printed text and handwriting content could be close or touch each other. In either the situation of OCR(Optical Character Recognition) or that of handwriting recognition tasks like signature verification [5] the redundant parts of printed text components will harm the performance of classifier with unexpected extent. To separate the printed text and handwriting text that are close spatially features associated with the size, stroke shape and spatial properties of printed text components should be extracted. Unfortunately not many works have been done on this in the literature [2].

In this paper we present a simple and practical model to realize the desired function in removing printed text from signature blocks. Motivated by the fact that even by local perceptual view people can easily recognize the printed text from the handwriting one, we construct a multi-layer model based on shape information, font size, and spatial position of disconnect components. Due to the fact that many researchers have developed powerful models in segmenting document content into respective indexed content we refrain our work on the cropped zone area where is indexed with handwriting text but overlapped with printed text that is regarded as noise.

2 Proposed Method

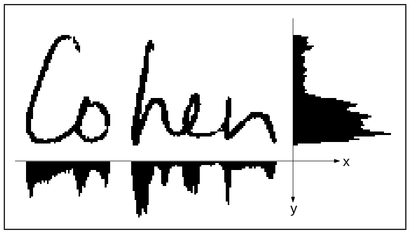
Existing document decomposition methods can divide page components into appropriate categories when the the document is clean and has non-overlapping fields. Our interest is in real-world business documents where cropped handwriting blocks include printed text. Such cropping can be performed manually, as in interactive forensic document examination, or produced by a page segmentation algorithm. The handwriting block, e.g., a rectangular region containing the signature, can include printed text that is irrelevant to handwriting analysis and needs to be filtered out(Fig.2).

To extract the discrete connected components a chaincode generation procedure should be applied on the questioned image [6]. By scanning the image from top to bottom and left to right, and then tracing the boundaries of connected foreground regions we can obtain the chaincode information which includes the size of bounding box for each connected components. With the aid of chaincode information we can implement the following procedure to realize the function of filtering printed text.

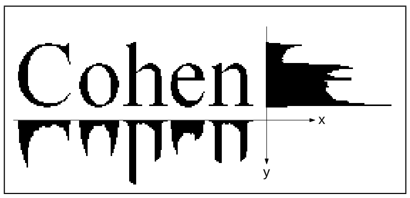
Step 1: Extract features and determine the candidate components to be removed. Observing that the font size of printed text should be consistent over the whole image, we extract two features, x_1, which is the perimeter of bounding box of individual component and is normalized by dividing over the maximum one, x_2, which is the ratio between height and width of bounding box of discrete components(Fig.3,4). The resolution of scanning does not affect the values since they are ratios. Height, width and perimeter are measured in terms of the number of pixels. For a skew-free image, the peaks of horizontal and vertical projection profiles of printed text components are expected to upright and tall(Fig.1). So we extract another two features, x_3 and x_4, which are the mean square values of projection profile's first derivative along y and x co-ordinates respectively.

$$\begin{aligned} x_3 &= \tfrac{1}{m-1}\textstyle\sum_{i=1}^{m-1}(Proj_y(i) - Proj_y(i+1))^2 \\ x_4 &= \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n-1}(Proj_x(i) - Proj_x(i+1))^2 \end{aligned}$$

where m and n are the height and width of text component.



(a)



(b)

Figure 1. The projection profiles. (a) The handwriting "Cohen" (b) The printed text "Cohen" using Times New Roman.

In the training process we manually separate the discrete printed text components from the handwriting components and extract the desired features. Fisher linear discriminant are chosen to find the optimum direction **w** in four-dimensional space to separate the two categories of components in sense of ratio of average inter-class distance and intra-class distance(Fig.5)(Only include the first two features to demonstrate in the two-dimensional surface).

The Fisher linear discriminant method employs that linear function $\mathbf{w}^T\mathbf{x}$ for which the criterion function

$$J(\mathbf{w}) = \frac{\|\tilde{m}_1 - \tilde{m}_2\|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \tag{1}$$

is maximum. Here $\tilde{m}_i$ are the means for projected samples of printed components and handwritten components respectively, and $\tilde{s}_i^2$ are the corresponding within-class scatters. The well known solution is

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{2}$$

Where $\mathbf{S}_W$ is called *within-class scatter matrix*.

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2 \tag{3}$$

and

$$\mathbf{S}_i = \sum_{\mathbf{x}\in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \tag{4}$$

See [7] for the details.

$y = \mathbf{w}^T\mathbf{x}$ which is the combined feature is further assumed to satisfy Gaussian distribution and all the disconnected components of the image are separated into two classes, suspected or unsuspected by the Bayesian approach. The discriminant function $g(y)$ is expressed as

$$g(y) = \frac{-(y-\tilde{m}_1)^2}{2\sigma_1^2} + \frac{(y-\tilde{m}_1)^2}{2\sigma_1^2} + ln\frac{\sigma_2}{\sigma_1} + ln\frac{P(w_1)}{P(w_2)} \tag{5}$$

When $g(y) > 0$ then class w_1 is chosen, otherwise class w_2 is chosen. Here w_1 and w_2 are indexes of printed text and handwritten text respectively. The inter-class variance σ and probability of appearance for the two classes, printed text and handwritten text are estimated from the training data set.

Thus after the first step we can determine the candidate printed text component set(Fig.6). All the components that will be removed should be from this set but the converse is not true.

Figure 2. Original mixed image.

Figure 3. Discrete connected components are in the boundary boxes.

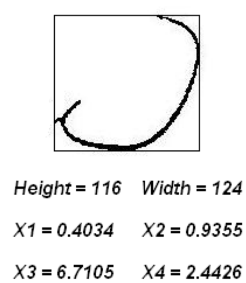


Figure 4. A discrete component piece with extracted features.

Step 2: In this step among the suspected components only the ones which are arranged in horizontal lines are erased from the original image. To be specific for every suspected component we compared the bounding box's y coordinate of the top and y coordinate of the bottom, if they are on the neighborhood of another suspected comexcept for this component's pixels it will be classified as printed text component and removed(Fig.8).

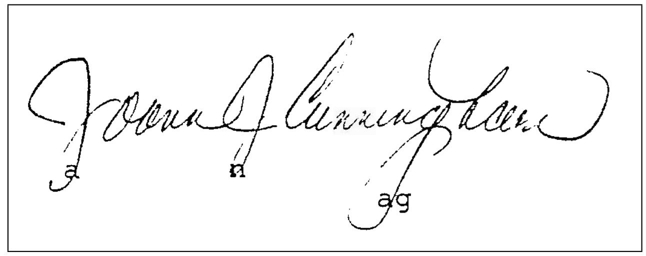


Figure 8. The final enhanced image.

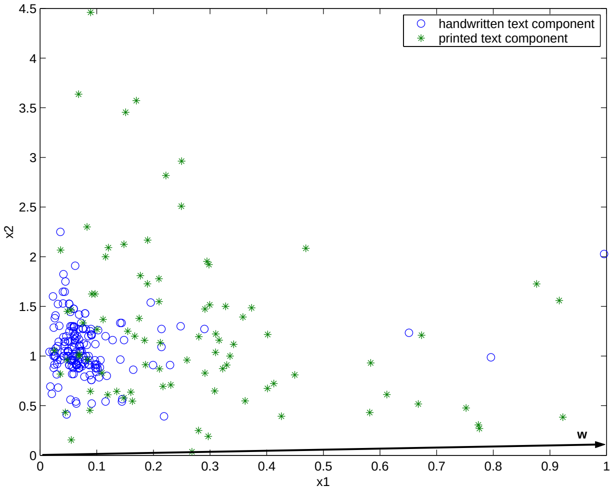


Figure 5. Feature distribution in two dimensional space constructed by x_1 and x_2. Optimum direction **w** is drawn.

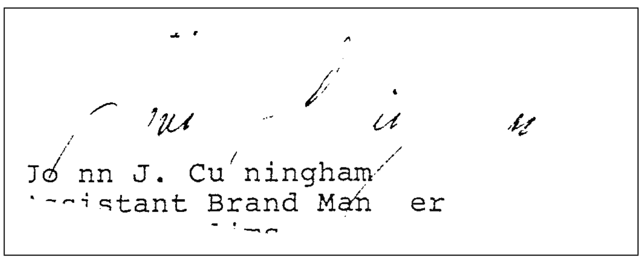


Figure 6. The candidates of printed text components.

ponent then it is determined as printed text component and removed from the original image(Fig.7).

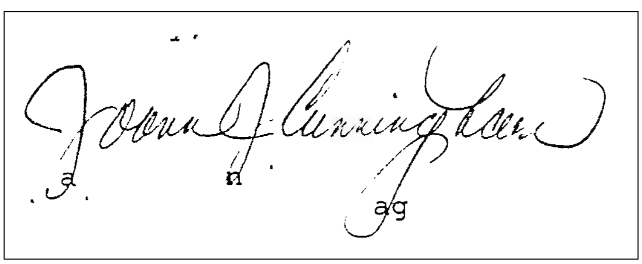


Figure 7. The processed image after step 2.

Step 3: Remove the far isolated components from the image. Though we can detect most printed text components by step 2 we take the risk ignoring some isolated printed text components because it is impossible to find a peer printed component in the same horizontal line. Further by doing this we can remove some trivial noise symbol like the punctuations which also have negative affect on the recognition. Here a window-based filtering is applied on every suspected components survived from step 2. The idea is to expand the bounding box of suspected component and find out if it is under the neighborhood of other stroke components by scanning the pixels inside the window. When there is no foreground pixel detected

3 Data Set

The data set that we use are the manually cropped signature images from the signed documents. There are totally 429 images in this data set. Each signature image consists of surrounding printed text, touching printed text and scanning noise. In the training process twenty images from the data set were randomly selected. These images were manually transformed into two versions: the one with only the handwriting components, and the one with the maximum size component and all the printed text(Fig.10). We kept the maximum component in the second version image because the feature x_2 is the perimeter of bounding box normalized by that of maximum component. From the former we can extract the 4 proposed shape features for handwriting text and from the latter we can have features for printed text. Thus we construct the training sets of two classes. The statistics for Fisher Linear Discriminant are listed below.

$$\begin{aligned}\mathbf{m}_1 &= (0.0843, 1.0782, 15.7823, 16.8234)^T\\ \mathbf{m}_2 &= (0.2726, 1.2794, 10.8627, 10.6608)^T\end{aligned}$$

$$\mathbf{S}_w = \begin{bmatrix} 5.9553 & -3.2405 & 262.3727 & 44.5444\\ -3.2405 & 71.7762 & -661.3069 & 726.8473\\ 262.3727 & -661.3069 & 54478 & 0.0185\\ 44.5444 & 726.8473 & 0.0185 & 3.8023 \end{bmatrix}$$

$$\mathbf{S}_w^{-1} = \begin{bmatrix} 0.2162 & 0.0039 & -0.0010 & -0.0003\\ 0.0039 & 0.0202 & 0.0002 & -0.0004\\ -0.0010 & 0.0002 & 0.0000 & -0.0000\\ -0.0003 & -0.0004 & -0.0000 & 0.0000 \end{bmatrix}$$

$$\begin{aligned}\mathbf{w} &= \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\\ &= (-0.0483, -0.0061, 0.0002, 0.0003)^T\end{aligned}$$

4 Experimental Results

The experiment of algorithm shows that visually over 85% of discrete printed text components in the 429 images have been removed and the handwritten text content has been well preserved. The remaining noise is contributed by the printed text overlapped with the signature piece which can not be removed in the connected component level (Fig.9). The performance evaluation of six images in Fig.9 is listed in Table 1, which is part of the statistics

Figure 9. Sample documents and their processed versions: Before(left column), After(right column).

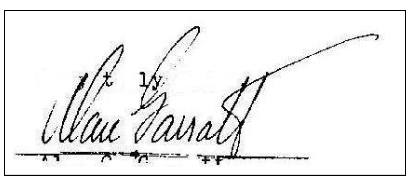
we got for all the testing images. We did the image retrival experiment [8] with the enhanced signatures but the performance does not improve much compared with the result using the original signatures. There are several possible reasons for that: one is that the data set we use is not big enough to indicate the efficiency of this algorithm and secondly the Gradient,Structural and Concavity features [9] we extracted from the images are not sensitive to small piece components like the printed text.

Table 1. Performance evaluation for images in Fig.9.

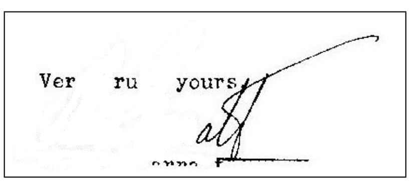
No.	A	B	C	D	E	F	G
1	57	46	1	6	97.56	62.5	87.72
2	41	29	2	0	93.55	100	95.12
3	111	71	0	4	100	90.91	96.4
4	36	25	0	0	100	100	100
5	41	33	1	1	96.97	87.5	95.12
6	29	22	1	0	95.65	100	96.55

A: Number of connected components
B: Number of detected printed text components
C: Number of error classified printed text components
D: Number of error classified handwritten components
E: E=1-C/[(B-D)+C] accuracy for printed text class
F: F=1-D/[(A-B-C)+D] accuracy for handwritten text class
G: G=1-(C+D)/A Overall accuracy

Note: Any component with more than 4 pixels appearing in the track of handwriting stroke should be counted as handwritten component. And it is similar for printed text components.



(a)



(b)

Figure 10. (a) Example in training set 1: handwriting image after removing the discrete printed text components, (b) Example in training set 2: image with handwriting components except for the maximum one being manually removed.

5 Discussion and Conclusion

We have presented a simple and useful image processing method aimed at separating printed text and handwriting text. The experiment proves that the selected features, the normalized size of bounding box around the pixel, and ratio between height and width of bounding box and the mean square value of projection profile's derivative are powerful in dricriminating these two classes. The spatial property of printed text component, that they are always located in a line, prevents the small piece handwriting component being removed in error. However to have the pure handwriting text it is necessary to recognize the printed text overlapped with handwriting text. To further improve performance the procedure must be taken to the stroke level, which will be the focus of our future work.

Acknowledgement

This work was supported in part by a grant from the Department of Energy through the Illinois Institute of Technology.

References

[1] Lawrence O'Gorman, "The Document Spectrum for Page Layout Analysis", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, November 1993, pp.1162-1173.

[2] Yefeng Zheng, Huiping Li and David Doermann, "Machine Printed Text and Handwriting Identification in Noisy Document Images", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, March 2004, pp.337-353.

[3] Anil K. Jain and Bin Yu, "Document Representation and Its Application to Page Decomposition", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, March 1998, pp.294-308.

[4] Hanchuan Peng, Fuhui Long and Zheru Chi, "Document Image Recognition Based on Template Matching of Component Block Projection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, September 2003, pp.1188-1192.

[5] Siyuan Chen, Sargur Srihari, "Combining One- and Two-Dimensional Approaches to Signature Verification",*Document Recognition and Retrieval XIII SPIE*, San Jose, CA, January 2006, pp.606701-1 to 606701-10.

[6] S. Madhvanath, G. Kim and V. Govindaraju, "Chaincode Contour Processing for Handwritten Word Recognitio", *IEEE Trans. on Pattern Analysis and Machine Intelligence* vol. 21, no. 9, September 1999, pp.928-932.

[7] Richard O. Duda, Peter E. Hart, David G. Stock, *Pattern Classification*, 2nd Edition, Johns Wiley & Sons, Inc., pp.117-121.

[8] Sargur Srihari, Shravya Shetty, Siyuan Chen, Harish Srinivasan and Chen Huang, "Document Image Retrieval using Signatures as Queries", *DIAL'06*, Lyon, France, April 2006, pp.198-203.

[9] J. T. Favata, G. Srikantan and S. N. Srihari, "Handprinted Character/Digit Recognition using a Multiple Resolution Philosophy", *Proc. International Workshop on Frontiers in Handwriting Recognition(IWFHR-94)*, 1994, pp.57-66.