The Segmentation and Identification of Handwriting in Noisy Document Images

Yefeng Zheng, Huiping Li, and David Doermann

Laboratory for Language and Media Processing
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742
{zhengyf, huiping, doermann}@cfar.umd.edu

Abstract. In this paper we present an approach to the problem of segmenting and identifying handwritten annotations in noisy document images. In many types of documents such as correspondence, it is not uncommon for handwritten annotations to be added as part of a note, correction, clarification, or instruction, or a signature to appear as an authentication mark. It is important to be able to segment and identify such handwriting so we can 1) locate, interpret and retrieve them efficiently in large document databases, and 2) use different algorithms for printed/handwritten text recognition and signature verification. Our approach consists of two processes: 1) a segmentation process, which divides the text into regions at an appropriate level (character, word, or zone), and 2) a classification process which identifies the segmented regions as handwritten. To determine the approximate region size where classification can be reliably performed, we conducted experiments at the character, word and zone level. We found that the reliable results can be achieved at the word level with a classification accuracy of 97.3%. The identified handwritten text is further grouped into zones and verified to reduce false alarms. Experiments show our approach is promising and robust.

1 Introduction

The ability to segment a document into functionally different parts has been an ongoing goal of document analysis research. In the case where the content is presented differently (in a different font, or as handwritten as opposed to machine printed), different analysis algorithms may be required for interpretation. In the case of handwritten annotations, such marks often indicate corrections, additions or other supplemental information that should be treated differently from the main or body content. We have found annotations of particular interest in the processing of correspondence and related business documents.

Previous work related to this problem has focused on distinguishing handwritten from machine-printed text with the assumption that the text region is available and/or segmented. The identification is typically performed at the text line [1, 2, 3, 4], word [5] or character level [6, 7]. At the line level, the printed text lines are typically arranged regularly, while the handwritten text lines are irregular. This characteristic is exploited by several researchers. Srihari et al. implemented a text line based approach and achieved the identification accuracy of 95% [1]. One advantage of the approach is

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 95–105, 2002.

it can be used in different scripts (Chinese, English etc.) with little or no modification. However, this feature is not available at the word level. Guo et al. proposed an approach based on the vertical projection profile of the word [5]. They used a Hidden Markov Model (HMM) as the classifier and achieved the identification accuracy of 97.2%. Although at the character level less information is available, humans can still identify the handwritten and printed characters easily, inspiring researchers to pursue classification at the character level. Kuhnke proposed a neural network-based approach with straightness and symmetry as features, and achieved an identification accuracy of 96.8% and 78.5% for the training and test sets respectively [7]. Zheng used a run-length histograms as features to identify handwritten and printed Chinese characters [6]. About 75% of the strokes are either horizontally or vertically straight in printed Chinese characters, but curved in handwritten characters. This distinctive characteristic is used to identify handwritten and printed Chinese characters. Based on the run-length histogram features, Zheng achieved the identification accuracy of 98%. However, the method cannot be extended to Latin character because the strokes of most Latin characters are curved.

Most of the previous research is focused on the identification problem with the assumption that the regions to be identified is already segmented or available. In practice, however, handwritten annotations are often mixed with printed text. The handwritten regions must be separated from the printed text first. Previous page segmentation algorithms can be classified into three categories: bottom-up, top-down and hybrid[8]. In a typical bottom-up approach[9], connected components are extracted then merged into words, lines and zones based on the spatial proximity. A top-down approach starts from the whole document and then splits it recursively into columns, zones, lines, words and characters[10]. No matter what segmentation method is used, the special consideration must be given to the size of the region being segmented. If the region is too small, the information contained in it may be not be sufficient for identification; if the region is too large the handwritten text may mix with the printed text in the same region. Figure 1 shows the relation between the accuracy of segmentation and the amount of information contained for identification at different region sizes. The experimental results presented later show the identification process is robust and reliable at the word level.

The diagram of our system is shown in Figure 2. After filtering the noise, we extract the connected components and merge them into words based on the spatial proximity. A trained Fisher classifier is then used to identify the handwritten and printed words. Finally, the identified handwritten words are merged into zones.

The rest of the paper is organized as follows: In Section 2 we present our segmentation and identification method. Section 3 describes the experimental results. Discussions and future work are provided in Section 4.

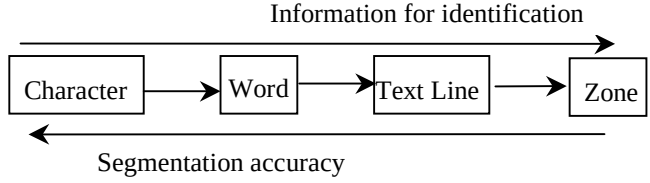


Fig. 1. The relation between the accuracy of segmentation and the amount of information contained for identification at different region sizes.

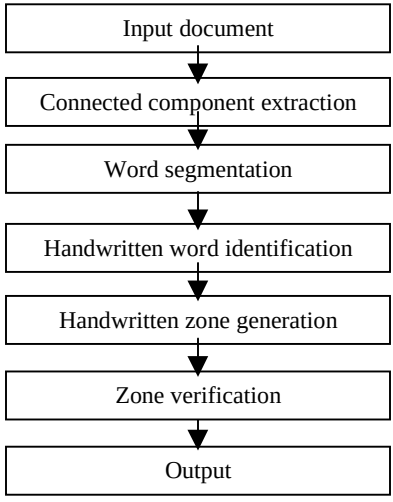


Fig. 2. The diagram of our system

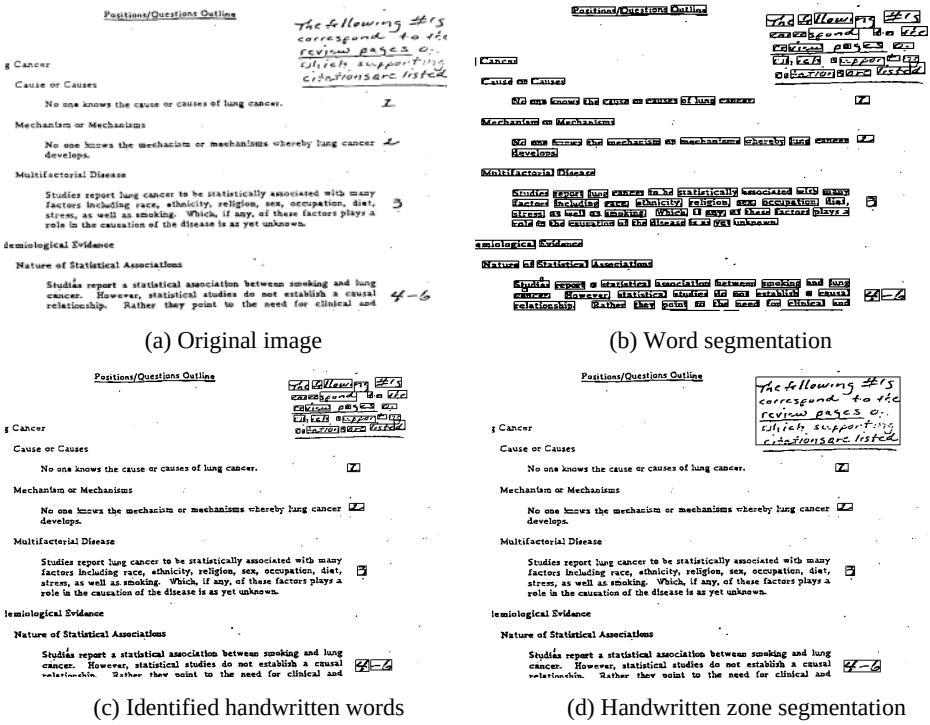
2 The Approach

In this section we will present our method for the segmentation and identification of handwritten annotations.

2.1 Word Segmentation

We use a bottom-up approach to segment the text into words. After the connected components are extracted from the document image, we estimate the average character size using a histogram of component heights [9]. We then group the neighboring connected components into words if they are spatially close in the horizontal direction. Sometimes handwritten annotations come in contact with or are very close to the printed text. When they are grouped into the same word due to the spatial proximity, we enforce the following rules: Two neighboring components $C1$ and $C2$ are merged only when they satisfy $\max(h_1, h_2) < 2\times\min(h_1, h_2)$, where h_1 and h_2 are the heights of $C1$ and $C2$ respectively.

Currently we assume the document has been de-skewed and the primary direction of the text line is horizontal. Figure 3 shows an example of the segmentation. Figure 3a is the original document with handwritten annotations and Figure 3b shows the segmentation result. We observe that sometimes spurious handwritten marks are not grouped into words due to the variability of the gap between characters. However, this will not affect the classification result significantly. After words are segmented, we perform the classification described in next section.



(a) Original image (b) Word segmentation

(c) Identified handwritten words (d) Handwritten zone segmentation

Fig. 3. Procedures of handwritten zone segmentation and identification

2.2 Handwritten Annotation Identification

In this section we present details of our handwritten/machine-printed text identification approach. After extracting structural and texture features, we use a Fisher classifier to map the extracted features to a value used to classify handwritten/printed text.

2.2.1 Feature Extraction

Structural Features

We use two groups of structural features. The first group is related to the physical size of the regions, including the width and height of the normalized region, the aspect ratio of the region and the density of the black pixels. The variance of the size distribution of the handwritten words is often larger than that of the printed words. We use a histogram technique to estimate the dominant font size, and then use the dominant font size to normalize the width and height of the region respectively. The aspect ratio of the region and the black pixel density are also used as features.

The second group contains features consisting of the average width and height of the connected components in the word, the aspect ratio of the connected components, the overlap rate of the connected components, and the variance of the projection pro-

file. In a handwritten region, the bounding boxes of the connected components tend to overlap with each other, resulting in a larger overlap area. The overlap areas are normalized by the total area of the region. In a machine-printed text region, the characters tend not to touch each other and therefore, the vertical projection profile has obvious valleys and peaks. The variance of the vertical projection is used to represent this characteristic.

Bi-level Co-occurrence

A co-occurrence histogram is the number of times a given pair of pixels occurs at a fixed distance and orientation. In the case of binary images, the possible occurrences are white-white, black-white, white-black and black-black at each distance and orientation. In our case, we are concerned primarily with the foreground. Since the white background region often accounts up to 80% of a document page, the occurrence frequency of white-white or white-black pixel pairs would always be much higher than that of black-black pairs. The statistics of black-black pairs carry most of the information. To eliminate the redundancy and reduce the effects of over-emphasizing the background, only black-black pairs are considered. Four different orientations (horizontal, vertical, major diagonal and minor diagonal) and four distance levels (1, 2, 4, 8 pixels) are used for identification (altogether 16 features). The details can be found in[11].

Bi-level 2×2-grams

The N×M-gram was introduced by Soffer in the context of image classification and retrieval [13]. We are using bi-level 2×2-grams at a hierarchy of distance from the origin. As described above, we first remove the dominant background (all the white background grams). We then scale each entry by multiplying the number of occurrence by a coefficient proportional to the number of black pixels in the 2×2-gram. The more black pixels, the larger the coefficient. In this work, we used $p^b(1-p)^{4-b}$, where p is the density of the image block, and b is the number of 1's in the 2×2-gram. We then normalize the entire vector of occurrences by dividing them by the sum of all occurrences. Four distances (1, 2, 4, 8 pixels) are used for identification (altogether 60 features). The details can be found in [11].

Pseudo Run Lengths

True run length counts are expensive to compute. We proposed a much faster method for computing pseudo run length statistics as features. The basic idea is we first down-sample the image to effectively preserve the low frequency components; the larger the down-sampling rate, the lower the frequency of the preserved components. By comparing the original signal with the down-sampled one, we can estimate the high frequency components that are present in the original signal. Down-sampling can be implemented efficiently using a look-up table. We do 1/2 down-sampling twice and get 16 features. The details can be found in [11].

Gabor Filters

Gabor filters can represent signals in both the frequency and time domains with minimum uncertainty [14] and have been widely used for texture analysis and segmentation [12]. Researchers found that it matches the mammal's visual system very well, which provides further evidence that we can use it in our segmentation tasks.

In spatial and frequent space, the two dimensional Gabor filter is defined as:

$$g(x,y) = \exp\left\{-\pi\left(\frac{x'^2}{\sigma_x^2} + \frac{y'^2}{\sigma_y^2}\right)\right\} \times \cos\{2\pi(u_0 x + v_0 y)\} \quad \textbf{(1)}$$

$$G(u,v) = 2\pi\sigma_x\sigma_y\left(\exp\left\{-\pi\left[(u'-u_0')^2\sigma_x^2 + (v'-v_0')^2\sigma_y^2\right]\right\} + \exp\left\{-\pi\left[(u'+u_0')^2\sigma_x^2 + (v'+v_0')^2\sigma_y^2\right]\right\}\right) \quad \textbf{(2)}$$

where $x' = -x\sin\theta + y\cos\vartheta$, $y' = -x\cos\theta - y\sin\theta$, $u' = u\sin\theta - v\cos\theta$, $v' = -u\cos\theta - v\sin\theta$, $u_0' = -u_0\sin\theta + v_0\cos\theta$, $v_0' = -u_0\cos\theta - v_0\sin\theta$, $u_0 = f\cos\theta$, $v_0 = f\sin\theta$. Here f and θ are two parameters, indicating the central frequency and orientation.

Suppose an original image is *I(x,y)*, then the filtered image *I'(x,y)* can be described as:

$$I'(x,y) = I(x,y) * g(x,y) \quad \textbf{(3)}$$

$$I'(u,v) = I(u,v)G(u,v) \quad \textbf{(4)}$$

It is very expensive, however, to calculate the filter in the spatial domain defined in Equation 3. Instead, we use an FFT to calculate it in the frequent domain. Let *I(u,v)* be the FFT of the original image and $G_k(u,v)$ be the frequency response of the k[th] Gabor filter. Then the frequency spectrum of filtered image $I_k'(u,v)$ equals to the product of *I(u,v)* and $G_k(u,v)$. The filtered image can be achieved by calculating the inverse FFT of $I_k'(u,v)$. For Gabor filters with different parameters, $G_k(u,v)$ is calculated and pre-stored. *I(u,v)* is calculated only once and shared among different Gabor filters. This can reduce the computation significantly. However, it is still expensive to calculate FFT when image regions are big. To reduce the computation further, we divide the whole region into several small blocks. Suppose the variance of the filtered small block using the k[th] filter is σ_i^k, then the feature of the filtered image is calculated as the weighted sum of the variance of each small block as described in Equation 5:

$$g_k = \frac{\sum_{i=1}^{N} w_i \sigma_i^k}{\sum_{i=1}^{N} w_i} \qquad k = 1,2,...16 \quad \textbf{(5)}$$

Where the weight w_i is the number of black pixels in the block.

For each orientation θ we can get a different filtered images and calculate the weighted variance (Equation 5) as a feature. In our experiments we let $\theta_k = k * \frac{180}{N}$, k = 1, 2, … N, with N = 16. Altogether there are 16 features.

2.2.2 Classification

We use Fisher classifier for classification. For a feature vector *X*, the Fisher classifier projects *X* onto one dimension *Y* in the direction *W*:

$$Y = W^T X \tag{6}$$

The Fisher criterion finds the optimal projection direction W_o by maximizing the ratio of the between-class scatter to the within-class scatter, which benefits the classification. Let S_w and S_b be within- and between-class scatter matrix respectively,

$$S_w = \sum_{k=1}^{K} \sum_{x \in \text{class } k} [(x - u_k)(x - u_k)^T] \tag{7}$$

$$S_b = \sum_{k=1}^{K} (u_k - u_0)(u_k - u_0)^T \tag{8}$$

$$u_0 = \frac{1}{K} \sum_{k=1}^{K} u_k \tag{9}$$

where u_k is the mean vector of the k[th] class, u_0 is the global mean vector and *K* is the number of the classes. The optimal projection direction is then the eigenvector of $S_w^{-1} S_b$, corresponding to the largest eigenvalue [15]. For two-class classification problems, we do not need to calculate the eigenvector of $S_w^{-1} S_b$. It is shown that the optimal projection direction is:

$$W_o = S_w^{-1}(u_1 - u_2) \tag{10}$$

Let y_1 and y_2 be the projection of two classes and $E[y_1]$ and $E[y_2]$ be the *mean* of y_1 and y_2. Suppose $E[y_1] > E[y_2]$, then the decision can be made as:

$$C(x) = \begin{cases} class1 & \text{If } y > (E[y1] + E[y2])/2 \\ class2 & \text{Otherwise} \end{cases} \tag{11}$$

It is shown that if the feature vector *X* is jointly Gaussian distributed, the Fisher classifier achieves optimal classification in a minimum classification error sense [15]. Figure 3c shows the result after the identification. Only the identified handwritten words are marked with rectangle boxes.

2.3 Handwritten Zone Generation

After identifying the handwritten words, we merge them into zones using the following rules:

1) Select the largest unmerged handwritten word as a seed.
2) Find the candidate word with the minimum distance to the seed.

3) If the minimum distance is smaller than a threshold (we choose four times of character width in the experiment), then group it with the seed.
4) Repeat Step 2 and 3 until no words can be grouped with the seed. The grouped region is marked as a handwritten zone.
5) Repeat Step 1 to 4 to generate all handwritten zones.

To reduce the false alarm, we run the identification process on the merged zones to verify further. Those zones with small confidence Figure 3d shows the extracted handwritten zone after merging words.

3 Experiments

3.1 Data Collection

We collected 318 documents containing handwritten annotations provided by the tobacco industry. Each handwritten zone and word is ground truthed for the evaluation purposes. However, the ground truth at the character level is expensive to achieve. Instead, we extract connected components inside the specified handwritten word as characters. Sometimes the handwritten characters touch each other so a connected component may contain several characters. It does not, however, affect the overall result significantly. All together we have 641 zones, 1504 words and 5177 characters in the specified handwritten zones. Since machine-printed characters outnumber handwritten characters, we randomly select roughly the same number of machine-printed characters, words and zones for experiments.

3.2 Identification of Handwritten/Machine-Printed Text

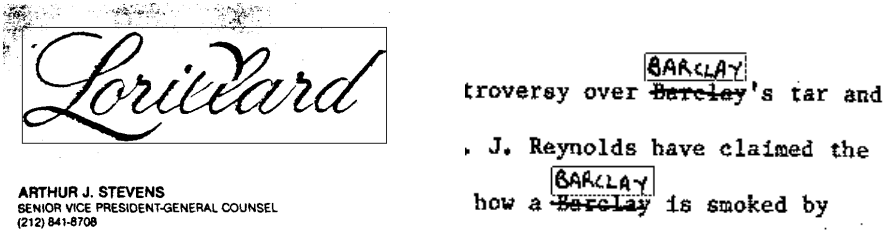
For the purpose of comparison, the experiment of handwritten/printed text identification is conducted at the character, word and zone levels. We use a N-fold cross validation technique to estimate the identification accuracy [15]. First, we divide the data into 10 groups. We then use one group as the test set and the remaining nine groups as the training set to conduct classification. This process is repeated ten times with a different group selected as the test set at each iteration. The average and variance of accuracy are shown in Table 1.

Table 1. Handwritten/printed text identification at different levels (in percentage)

	Structural features	Bi-level co-occurrence	Bi-level 2×2-grams	Pseudo run lengths	Gabor filter	All features
Character	84.4±1.0	83.4±0.5	87.6±1.1	84.4±0.8	86.6±0.8	93.0±0.6
Word	95.7±0.5	87.5±1.5	94.3±1.2	89.8±1.5	95.0±1.3	97.3±0.5
Zone	89.7±2.4	88.1±2.8	94.0±2.7	91.0±1.8	94.9±1.9	96.8±1.6

From Table 1, we can see the identification achieves the best result at the word level with an accuracy of 97.3%, which provides further evidence that it is appropriate

for us to segment the document at the word level. For all three levels, it shows the classification with all features achieves better result than with a single group of features.



(a). A printed zone is falsely identified as a handwritten zone due to the font

(b) Two handwritten zones are falsely identified as printed

Fig. 4. Identification errors

It may be surprising that the identification accuracy at the zone level is not as good as at the word level. We observed most identification errors at the zone level occur in the small zones containing only one or two words. Therefore, actually no more information can be used at the zone level than at the word level. Large zones containing more words can be identified more reliably, but the small zones containing fewer words are more error-prone. It is more reasonable to evaluate the zone level identification accuracy by considering the number of words in the zone as weights.

Figure 4a shows an identification error where a printed text zone is identified as a handwritten one because the font is so similar to handwriting. Figure 4b shows an example that two handwritten zones are falsely identified as printed.

3.3 Experiments on Handwriting Segmentation and Identification

We tested our algorithm on 318 documents. Figures 5 and 6 show some examples of segmented results. The identification error rate at the word level is between 2-3%. For a typical document, there are about 200 printed words. Therefore between 4 and 6 printed text zones will be identified as handwritten zones. Another type of error is that some noise is segmented and identified as handwritten zones. The problem occurs when the document is extremely noisy. Figure 7 shows the segmentation and identification result for an extremely noisy document image. Although our system identifies all the handwritten text, other regions (logo, noise, etc) are also identified as handwritten text. We are actively investigating more features and classifiers to improve the result.

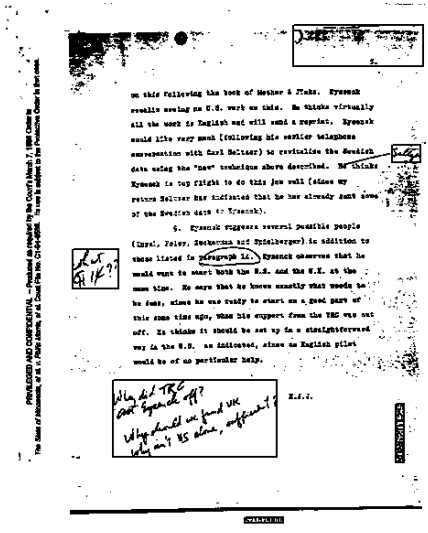


Fig. 5. One example of handwritten zone segmentation and identification

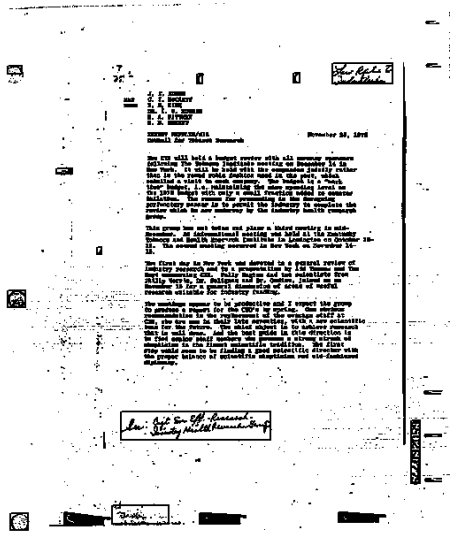
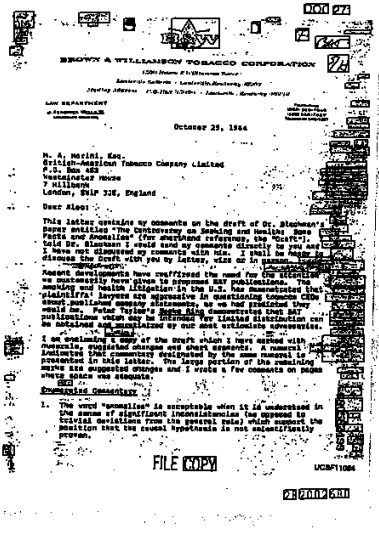
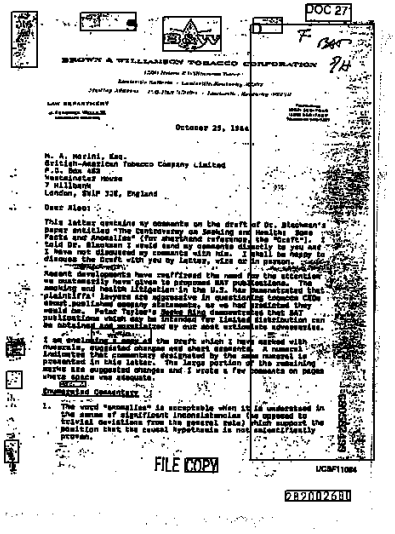


Fig. 6. Another example of handwritten zone segmentation and identification



(a) Handwritten word identification result



(b) Handwritten zone segmentation result

Fig. 7. A challenge case on an extremely noisy document image. Although all of the handwritten are correctly identified, some of the noise regions are incorrectly segmented as handwritten

4 Conclusions and Future Work

We have presented an algorithm to segment and identify handwritten/printed text in document images. Our approach consists of two processes: 1) a segmentation process, which segments the text into regions at the word level, and 2) a classification process which identifies the segmented handwritten regions. The experimental results show our method is promising.

We are actively extending this work in three directions. First, our current method filters the noise by the size, which is not robust enough when document is extremely noisy. We need to explore more robust features and classifier to identify handwritten/printed text and noise well. Second, we are developing a scheme to use contextual information to further increase identification accuracy. And at last we will quantitatively evaluate the final segmentation and identification result.

References

1. S. N. Srihari, Y. C. Shim and V. Ramanaprasad. A system to read names and address on tax forms. *Technical Report CEDAR-TR-94-2*, CEDAR, SUNY, Buffalo, 1994
2. K. C. Fan, L. S. Wang and Y. T. Tu. Classification of machine-printed and handwritten texts using character block layout variance. *Pattern Recognition, 31(9)*, pages 1275–1284, 1998
3. V. Pal and B. B. Chaudhuri. Machine-printed and handwritten text lines identification. *Pattern Recognition Letters*, 22, pages 431–441, 2001
4. J. Fanke and M. Oberlander. Writing style detection by statistical combination of classifier in form reader applications. In *Proc. of the 2nd Inter. Conf. On Document Analysis & Recognition*, pages 581–584, 1993
5. J. K. Guo and M. Y. Ma. Separating handwritten material from machine printed text using hidden Markov models. In *Proc. of the 6th Inter. Conf. On Document Analysis & Recognition*, pages 439–443, 2001
6. Y. Zheng, C. Liu and X. Ding. Single character type identification. In *Proc. of SPIE Vol. 4670, Document Recognition & Retrieval IX*, pages 49–56, 2001
7. K. Kuhnke, L. Simoncini and Zs. M. Kovacs-V. A system for machine-written and handwritten character distinction. In *Proc. of the 3rd Inter. Conf. On Document Analysis & Recognition*, pages 811–814, 1995
8. S. Mao and T. Kanungo. Empirical performance evaluation methodology and its application to page segmentation algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(3), pages 242–256, 2001
9. L. O'Gorman. The document spectrum for page layout analysis. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 15(11), pages 1162–1173, 1993
10. G. Nagy, S. Seth and S. Stoddard. Document analysis with an expert system. *Pattern Recognition in Practice II*, Elsevier Science, pages 149–155, 1984
11. D. Doermann and J. Liang. Binary document image using similarity multiple texture features. In *Proc. of Symposium on Document Image Understanding Technology*, pages 181–193, 2001
12. A. K. Jain and S. Bhattacharjee. Text segmentation using Gabor filters for automatic document processing. *Machine Vision Application*, 5, pages 169–184, 1992
13. A. Soffer. Image categorization using texture features. In *Proc. of the 4th Inter. Conf. on Document Analysis & Recognition*, pages 233–237, 1997
14. D. Gabor. Theory of communication. *J. Inst. Elect. Engr.* 93, pages 429–459, 1946
15. K. Fukunaga. Introduction to statistical pattern recognition. Second edition, Academic Press Inc. 1990