Signature Segmentation from Machine Printed Documents using Conditional Random Field

Ranju Mandal
Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata-108, India
ranjumandal@gmail.com

Partha Pratim Roy
Laboratoire d'Informatique
Université François Rabelais
Tours, France
partha.roy@univ-tours.fr

Umapada Pal
Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata-108, India
umapada@isical.ac.in

***Abstract*—Automatic separation of signatures from a document page involves difficult challenges due to the free-flow nature of handwriting, overlapping/touching of signature parts with printed text, noise, etc. In this paper, we have proposed a novel approach for the segmentation of signatures from machine printed signed documents. The algorithm first locates the signature block in the document using word level feature extraction. Next, the signature strokes that touch or overlap with the printed texts are separated. A stroke level classification is then performed using skeleton analysis to separate the overlapping strokes of printed text from the signature. Gradient based features and Support Vector Machine (SVM) are used in our scheme. Finally, a Conditional Random Field (CRF) model energy minimization concept based on approximated labeling by graph cut is applied to label the strokes as "signature" or "printed text" for accurate segmentation of signatures. Signature segmentation experiment is performed in "tobacco" dataset[1] and we have obtained encouraging results.**

Keywords- Signature segmentation, Printed/handwritten text separation, Signature verification, CRF

I. INTRODUCTION

Identification of handwritten annotations and signatures made on machine printed documents is important for document interpretation. The aim is to segment such mixed documents into two layers: a layer assumed to contain printed text and other layer contains the handwritten parts. Such segmentation problem has received a great deal of attention in the literature because of the different processing approaches for printed and handwritten texts. The objective is to apply respective techniques on the printed and handwritten parts.

Signature is often examined by forensic document analysis and the banking and finance industry to restrict frauds. Thus, signature authentication is being carried out to verify signature. Many research works are going on for automatic online/offline signature verification and recognition [2, 6, 8]. However, these processes assume that the signatures are isolated and they do not touch/overlap with other text in the document.

A machine printed document which contains a signature, there may be some printed texts that may touch and/or overlap the signature. We have shown an example of such signed document in Fig.1. It is to be noted that the signature strokes are overlapped/touched with printed text characters in many places. Some of the overlapped regions are marked (in red) by a rectangular box in the zoomed version. Proper segmentation of such touching signature is needed before applying the methods for signature verification and recognition.

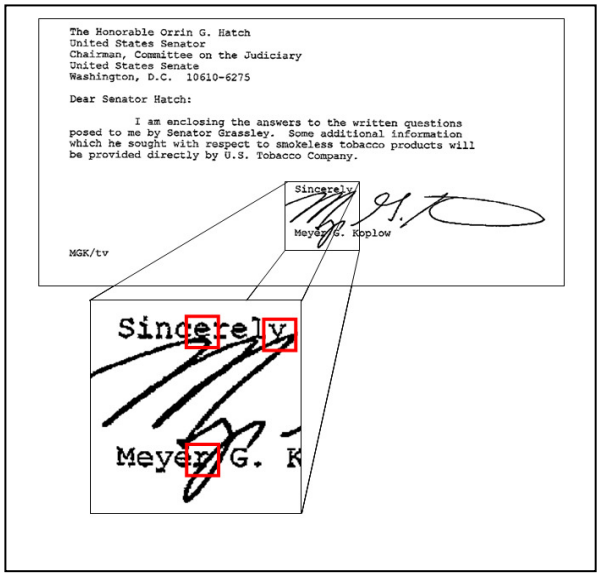


Figure 1. . Sample document of overlapped signature on a machine printed text. A zoomed version of some touching portions in the signature is also shown below the document.

Although many algorithms developed are for detection of handwritten annotation and signature verification and recognition, much effort is not given towards signature segmentation from a document page. Different types of segmentation and feature extraction methods, and various classification models have been proposed inthe literature. Shetty et al.[14] proposed a method for automatic labeling of scanned document containing handwriting and machine printed texts using Conditional Random Field (CRF). Farooq et al. [4] have proposed Gabor filters for feature extraction and an Expectation Maximization (EM) based probabilistic neural network for classification. Guo and Ma [3] used Hidden Markov Models (HMM) based classification for handwritten annotation separation from printed document.

[1] http://legacy.library.ucsf.edu/

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.236

Peng et al.[1] have used a modified K-Means clustering algorithm for classification at an initial stage and then Markov Random Field (MRF) have been used for relabeling. In another work, overlapped texts are segmented by shape context based aggregation and MRF [5].

Most of the earlier work done on printed text and handwritten annotation separation are based on word level classification. To deal with overlapping/touching part of the signature strokes are analyzed in our scheme and only a few works deal with this problem. In this paper, we focused on the segmentation of signature from a document by eliminating the printed text that touch or overlap the signature.

A block diagram of our proposed approach is shown in Fig.2. A two-stage approach has been proposed here for signature segmentation. In the first stage, printed and handwritten word blocks are separated. To do so, at first, words blocks are extracted from a document image and the signature in these word blocks are detected using block level feature analysis. A signature block may contain some printed characters because of the overlapping of signature with the printed text. Thus, in the second stage, a stroke level segmentation and classification are performed using skeleton analysis to separate the overlapping printed text from signatures. We have used 400 dimensional gradient based features and Support Vector Machine (SVM) for classification in both blocks and stroke level classification. Finally, a Conditional Random Field (CRF) model energy minimization concept based on approximated labeling by graph cut is applied for final classification of printed and signature strokes to obtain more accurate segmentation of signature portions from printed text.

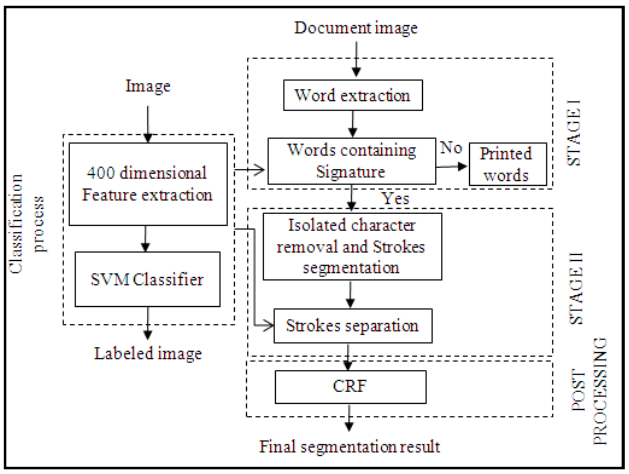


Figure 2. Block diagram of the proposed system

The organization of the rest of the paper is as follows: In Section II, we have discussed briefly the feature extraction and the classifier used for signature detection and stroke classification. The proposed methodology of signature segmentation is detailed in Section III. We have demonstrated the experimental results and analyzed the performance in Section IV. Finally conclusion and future work are presented in Section V.

II. Feature Extraction and Classification

A signature generally consists of some large strokes n compare to the strokes of the printed text. So, this distinct feature of signature is very important to get the difference of signature from printed strokes. After detecting a signature block we followed few steps to segment it into strokes. We compute 400 dimensional gradient based features in the both levels of our scheme and the feature extraction technique is described below.

A. 400 dimensional gradient feature

To obtain 400 dimensional feature [7] the following steps are applied. At first, size normalization of the input binary image is done. Here we normalize the image into 126x126 pixels. The input binary image is then converted into a gray-scale image by applying a 2x2 mean filtering 5 times. The gray-scale image is normalized next so that the mean gray scale becomes zero with maximum value 1. The normalized image is then segmented into 9X9 blocks. A robust filter is then applied on the image to obtain gradient image. The arc tangent of the gradient (strength of gradient) is quantized into 16 directions (an interval of 22.5°) and the strength of the gradient is accumulated with each of the quantized direction. By strength of Gradient f(x, y) we mean $f(x, y)=\sqrt{(\Delta u)^2+(\Delta v)^2}$ and by direction of gradient $(\theta(x,y))$ we mean $(\theta(x,y)) = tan^{-1}\frac{\Delta u}{\Delta v}$, here $\Delta u = g(x+1,y+1) - g(x,y)$, $\Delta v = g(x+1,y) - g(x,y+1)$ and $g(x,y)$ is a gray scale value at an (x, y) point.

Histograms of the values of 16 quantized directions are computed in each of 9x9 blocks. Finally, 9x9 blocks are down sampled into 5x5 by a Gaussian filter. Thus, we get 5x5x16 = 400 dimensional feature.

B. Classifier Details

In our experiments, we have used a Support Vector Machine (SVM) as classifier. The SVM is defined for two-class problem and it looks for the optimal hyper plane which maximizes the distance, the margin, between the nearest examples of both classes, named support vectors (SVs). Given a training database of M data: $\{x_m \mid m=1,...,M\}$, the linear SVM classifier is then defined as:

$$f(x) = \sum_j \alpha_j x_j \cdot x + b$$

Where $\{x_j\}$ are the set of support vectors and the parameters α_j and b has been determined by solving a quadratic problem [11]. The linear SVM can be extended to various non-linear variants, details can be found in [11]. In our experiments Gaussian kernel SVM outperformed other non-linear SVM kernels, hence we are reporting our recognition results based on Gaussian kernel only. The Gaussian kernel is of the form:

$$[k(x, y) = \exp(-\frac{\| x-y \|^2}{2\sigma^2})]$$

.

We noticed that Gaussian kernel gave higher accuracy when the value of its gamma parameter is 12.00 and the penalty multiplier parameter is set to 10.

III. PROPOSED APPROACH

A histogram-based Otsu binarization method is applied to convert the document image into two-tone images. The binary image may contain some spurious noise pixels and irregularities on the boundary of the characters, leading to undesired effects on the system. We removed such small noise components and smoothed the rest of the image for signature segmentation.

A. Signature detection

The binarized document image is segmented into words based on the inter-character spacing in words. To do so, we have performed a morphological dilation operation to segment the words. The size of the structuring element for dilation is chosen as square element of 5x5. Then, a connected component labeling is applied to find the bounding box of the word patches in the dilated image. Based on the positional information of the bounding box of the word patches, the respective positions of the words are then segmented from original document. Next, we compute 400 dimensional gradient based feature of each of the segmented words and classify them as printed-text block or signature block. SVM classifier is used for this purpose. Different word examples of printed text and handwritten signature parts have been trained for this purpose. See Fig.3, where the word blocks are classified as signature and non-signature blocks using the above approach. Rectangular box marked by red box indicates the printed text (non-signature block) and the green box indicates signature block.

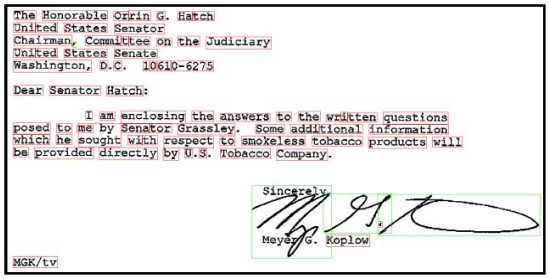
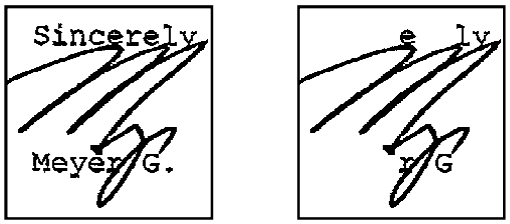


Figure 3. Document after detection of signature block

There may exist some isolated printed text in the signature word block due to word extraction after morphological operation. The printed isolated text characters that are very near to signature may be included with the signature block. These isolated characters are eliminated from signature block by checking the neighboring word information. We consider total 8 neighbor printed words and estimate the height and width information of text characters. The neighbor word blocks are decided using boundary growing algorithm. Boundary growing is done by expanding the boundary box of the word outwards iteratively by one pixel (8 pixel neighbour configurations).

The words that are touched first during boundary growing are selected as nearest neighbor word. We compute the average height and width information from these neighbor word blocks. The connected components of the estimated size in the signature block are checked for removal. Next, we check the local linearity of these isolated components by Hough Transform. If the size of a component is less than the estimated printed text character size and they follow local linearity, these text components are eliminated. See Fig.4(b), where we have shown the removal of isolated text component from a signature block of Fig.4(a).



(a) (b)

Figure 4. Isolated characters in a signature block of (a) are removed in (b).

B. Stroke based segmentation

The isolated printed text characters that are included in the signature block are removed. But, there may be some text characters that touch/overlap the signature. We have performed a stroke based segmentation analysis to remove these touching printed text characters.

To do so, at first the signature blocks are segmented into their constituent strokes. The decomposition of these strokes is performed by analyzing the thinned image of the signature blocks. For this purpose, a rotation invariant rule-based thinning algorithm [10] is applied to each of the signature block and junction points are searched. The junction points in this thinned image are found by detecting the pixel locations having 3 or more neighbors. In Fig.5, we show the junction points (red dots) of signature blocks obtained from thinned image. Next, each signature block is segmented into small strokes at the junction points. Due to spurious effect of thinning, sometimes the strokes can be over-segmented. To avoid this, the image is smoothed before thinning process.

After decomposition of these signature blocks, a stroke level classification is performed. 400 dimensional gradient based feature and SVM classifier are applied for separation of printed text strokes and signature strokes.

Figure 5. Detection of junction points in the signature blocks

C. Refinement by Conditional Random Field

In order to achieve more accurate signature segmentation, new constraints are introduced to reduce misclassifications that occur near the segmentation strokes. We formulate each strokes found at junction points into a node. Given the stroke set $X = (x_1, x_2,...)$, on which a 2D undirected graph is constructed, the objective is to find the best stroke label $Y = (y_1, y_2,...)$ to minimize the total graph energy E. Let $G(S, E)$ be the adjacency graph of segmented strokes $s_i \in S$ in the signature block. E is the set of edges formed between pairs of adjacent strokes (s_i, s_j) in the image.

Conditional random fields [12] provide a natural way to incorporate such constraints by including them in the pairwise edge potential of the model. Let $P(\mathbf{c}|G)$ be the conditional probability of the set of class label assignments c given the adjacency graph $G(S,E)$ and a weight w:

$$-\log(P(\mathbf{c}|G;w)) = \sum_{s_i \in S} \Psi(c_i|s_i) + w \sum_{(s_i,s_j)\in E} \Phi(c_i, c_j|s_i, s_j)$$

The unary potentials Ψ are defined by the probability classification score provided by SVM classifier for each stroke:

$$\Psi(c_i|s_i) = -\log(P(c_i|s_i))$$

and the pair wise edge potentials Φ are as follows.

$$\Phi(c_i, c_j|s_i, s_j) = \left(\frac{1}{1+L(s_i,s_j)}\right)$$

where, $L(s_i, s_j)$ is the distance between CG (centre of gravity) of strokes s_i and s_j. Weight w represents the trade-off between spatial distance and stroke confidence in the classification. The estimation of w is done by cross validation on the training data. When energy function parameters are learned, multi-label graph optimization technique: graph cuts (α-expansion) algorithm [13] is applied to find the best label Y of strokes to minimize the total energy since it can achieve approximate optimal results. See Fig.6, where we have shown the final segmentation result of a signature block shown in Fig.4(b). The signature and printed text characters strokes are shown in two different colours: red and blue, respectively for better visibility of segmentation results.

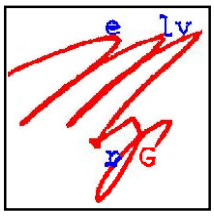


Figure 6. Final result of signature stroke segmentation in a touching signature block after applying CRF.

IV. Result and Disscussion

To the best of our knowledge, there exists no standard database to evaluate signature segmentation methods. For our experiment, we have used the dataset of "tobacco" industrial archives. The data set consisting of machine printed documents along with signature in many pages. We have used 105 signed machine printed document for the performance evaluation. The documents are written in English and the signatures on these documents also contain English text characters.

For quantitative performance of the system, we use common ratio of precision (P) and recall (R) for evaluation of stroke classification. Depending on the ground truth of the data each stroke is tested its belongingness to a signature or not. The precision measures the quality of the labeling in terms of the ability of the system to include only signature strokes. Whereas the recall measures the effectiveness of the system in extracting relevant signature strokes.

A. Training Set

To train our classifier for detecting a signature blocks on printed document, we have used 3080 signature from GDPS signature dataset [9] and 7684 English words as training data sets. English words are extracted from different types of printed documents like books, daily newspaper, official documents, magazine, journal etc. Also, to train the classifier for separating signature and printed strokes at stroke-level classification, we have used 2884 signature strokes and 10267 printed strokes. The signature strokes are extracted from 300 signatures and printed strokes are extracted from machine printed characters from the tobacco dataset.

B. Results

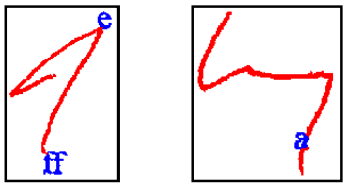
Block-level results: For the experiment, we have tested a total of 16743 patches (16303 printed word patches and 440 signature patches) and we have achieved an overall accuracy of 98.56%. We also noted that our method has successfully detected 432 signature patches out of 440 signature patches. From the experiment we noted that main reason of most of such errors is due to the smaller size of character components in the signatures.

Stroke-level results: In the stroke level classification, precision and recall results have been computed for separation of printed strokes and handwritten strokes from signature block. For this experiment we considered 1012 strokes. We produced two sets of results. Initially, strokes have been separated using Support Vector Machine and finally, CRF based post-processing model has been applied to improve the result. Table 1 shows these two sets of results of our experiment. It is to be noted that, we have obtained better stroke segmentation results using CRF.

TABLE I. Stroke Level Segmentation Result

Type	SVM		Using CRF	
	Precision	Recall	Precision	Recall
Printed stroke	91.86	90.61	93.1	90.82
Signature stroke	83.88	80.36	85.72	81.86
Overall	87.87	85.48	89.41	86.34

To have a qualitative idea of the stroke segmentation result, we have shown an example in Fig.6. The figure shows that, the printed characters ‘e’, ‘l’ and ‘y’ are successfully segmented from the signature stroke. The character ‘G’ could not be separated because the junction point between ‘G’ and signature stroke was not identified during stroke segmentation analysis. The character ‘r’ that was overlapped with the signature block is also separated but the overlapped portion of the signature stroke is labeled as printed text. We show two more segmentation results in Fig.7. The signature strokes are efficiently separated from these signature blocks by our proposed approach.



(a) (b)

Figure 7. Signature stroke segmentation result from different signature blocks.

Most of the errors of this stage are because of improper segmentation of strokes. Few strokes are over-segmented and some errors are encountered. On the other side, few overlapped printed strokes have not been segmented properly from the handwritten strokes and they have been miss-classified as handwritten strokes.

V. CONCLUSION

Signature detection and extraction from a document is an important task before feeding them for signature verification and recognition. In this paper, we have proposed a novel approach for detection and extraction of signature from machine printed documents. A stroke segmentation and classification based method is applied to extract the actual signature strokes eliminating overlapping and touching printed characters from them. We have used SVM to classify the signature strokes from machine printed strokes. Finally, the relationship of neighbor strokes are exploited using CRF based model to improve the stroke classification result. The experimental results demonstrate that the performance of our system is encouraging. There are scopes for improvements using this approach by extending the investigation to more accurate segmentation and classification.

REFERENCES

[1] X. Peng, S Setlur, V Govindaraju, R Sitaram and K Bhuvanagiri, “Markov Random Field Based Text Identification from Annotated Machine Printed Documents”, In Proc. 10th ICDAR, pp.431-435, 2009.

[2] M. A. U. Khan, M. K. K. Niazi and M. A. Khan, “Velocity-Image Model for Online Signature Verification", IEEE Transactions on Image Processing, pp.3540-3549, vol. 15, No. 11, 2006.

[3] J.K. Guo and M.Y. Ma, “Separating Handwritten Material from Machine Printed Text Using Hidden Markov Models”, In Proc. 6th ICDAR, pp.439-443, 2001.

[4] F. Farooq, K. Sridharan and V. Govindaraju, “ Identifying handwritten text in mixed documents”, Proc. International Conference on Pattern Recogniton, pp.1-4, 2006.

[5] X. Peng, S Setlur, V Govindaraju and Ramachandrula Sitaram, “Overlapped Text Segmentation Using Markov Random Field and Aggregation”, Proc. International Workshop on Document Analysis System , pp.129-134, 2010.

[6] J. F. V. Bonilla, M. A. Ferrer-Ballester, C. M. T. González, J. B. Alonso: Off-line signature verification based on grey level information using texture features. Pattern Recognition, vol. 44 no. 2, pp. 375-385, 2011

[7] U. Pal, N. Sharma, T. Wakabayashi, F. Kimura, “Handwritten Numeral Recognition of Six Popular Indian Scripts”. In Proc. 9th ICDAR , pp. 749-753, 2007.

[8] J.P.Swanepoel and J. Coetzer, “Off-line Signature Verification Using Flexible Grid Features and Classifier Fusion”, In Proc. ICFHR, pp. 297-302, 2010.

[9] M. Blumenstein, Miguel A. Ferrer, J.F. Vargas, “The 4NSigComp2010 off-line signature verification competition: Scenario 2”, In Proc. ICFHR, pp. 721-726, 2010.

[10] M.Ahmed and R. Ward, “A Rotation Invariant Rule-Based Thinning Algorithm for Character Recognition”, IEEE Transactions on PAMI,vol. 24, no. 12, pp.1672-1678, 2002.

[11] V.Vapnik, “The Nature of Statistical Learning Theory”, Springer Verlang, 1995.

[12] Y. Boykov and V. Kolmogorov, “An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision”, IEEE Transaction on PAMI, vol. 26, no. 9, pp. 1124-1137, 2004.

[13] V. Kolmogorov and R. Zabih, “What energy functions can be minimized via graph cuts?”, IEEE Transactions on PAMI, vol. 26, no. 2, pp. 147-159, 2004.

[14] Shravya Shetty, Harish Srinivasan and Sargur Srihari, “Segmentation and Labeling of Documents using Conditional Random Fields”, In Proc. Document Recognition and Retrieval IV, Proceedings of SPIE, vol.6500U,pp.1-11,2007.