Farsi/Arabic Handwritten from Machine-printed Words Discrimination

Saeed Mozaffari
Electrical and Computer Engineering Department
Semnan University
Semnan, Iran
mozaffari@semnan.ac.ir

Parnia Bahar
Electrical and Computer Engineering Department
Semnan University
Semnan, Iran
parniabahar@gmail.com

Abstract— **Separating handwritten texts from machine-printed materials is a desirable task towards a general document analysis system. In this paper, we proposed a simple and effective method to discriminate handwritten from machine-printed words in Farsi/Arabic documents. After finding word blocks, three different feature sets were extracted. They include two well-established features, previously used for Latin handwritten from machine-printed text separation, and a new feature, called baseline profile. Then, extracted features were combined together to obtain a feature vector with 34 elements. SVM and KNN classifiers were utilized to separate handwritten and machine-printed words. To evaluate the proposed method, some special forms, designed for word separation, were used. Experimental results show that our system differentiates between handwritten and machine-printed words with the overall accuracy of 97.1%.**

Keywords- *Farsi/Arabic Document Analysis*; *handwritten from machine-printed discrimination.*

I. INTRODUCTION

Comprehensive document analysis systems of the future should be able to handle documents with variety of information such as assortments of different language texts; combinations of figure, tables, and text; and mixtures of handwritten and machine-printed words. Dividing a document into handwritten and machine-printed areas is a useful task in many applications like bank check processing systems, automatic postal survives, and daily form analysis.

Usually, handwriting parts are used in conjunction with machine-printed documents to add corrections, additions, or other supplemental information. Since most of OCR systems need different methods to deal with handwritten and machine-printed documents, these parts should be separated from each other. Furthermore, in many applications, only one of the handwritten or machine-printed parts is of our interest. For example, in bank check processing, only handwritten fields have information, while in official documents retrieval, handwritten parts are considered as unnecessary information.

Previous work on this subject concerns the classification of text on the line-level, word-level or character-level, for Latin, non-Latin, or bilingual documents. Zheng et al. [1] perform text identification in noisy documents with comparative results for all levels. Fan et al. [2] detect handwriting by using structural characteristics for Chinese and English. Pal et al. [3] classify Indian scripts. Ma et al. [4] localize non-Latin script in Latin documents. Kavallieratou et al. [5] proposed a trainable approach with structural features to process Latin texts.

During past decade, we are witnessing a trend towards Arabic document analysis. This is mainly due to the large population of Arab world and the same characters set of other languages such as Farsi (Persian) and Urdu [6]. In this paper, we present a new method for Farsi/Arabic handwritten and machine-printed words separation. We also proposed a new feature, which is based on Farsi/Arabic language property, for document discrimination.

II. PROPOSED METHOD

Block diagram of the proposed method is shown in Figure 1. It mainly includes word blocking, feature extraction, and classification. Details of each section are presented here.

A. Database

For training and testing our system, some special forms are used in which handwritten and machine-printed parts are separated from each other by an indicator line at the middle of each form (figure 2). These forms are filled by several authors with different ages and educational backgrounds. Then, the collected forms are scanned in gray-level format with resolution of 300dpi [7].

B. Word Blocking

Handwritten from machine-printed texts discrimination can be performed at different levels. However, previous efforts demonstrate that texts separation at word-level outperforms line-level, and character-level [8]. So, in this paper, after finding connected components (CCs), they are combined together based on the relative distances between adjacent CCs to build words bounding boxes.

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.202

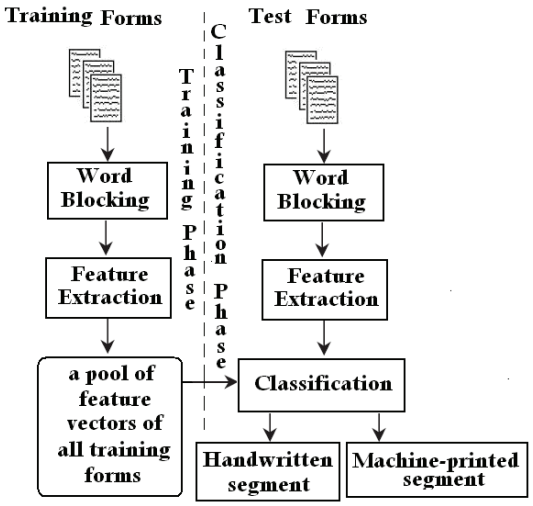


Figure 1. The proposed system overview

Unlike Latin language, Farsi/Arabic words can be composed of subwords, separated by small blanks. This property makes word blocking more difficult for Farsi/Arabic. Variety of handwriting styles in Farsi/Arabic documents, especially in complementary parts of some characters, and distances between adjacent subwords, makes word blocking a challenging task. So, applying the same word blocking algorithm to both handwritten and machine-printed documents led to different results. However, we make use of this diversity to differentiate handwritten and machine-printed word blocks in feature extraction step.

Figure 2. One special form for data collection.

First, all CCs are extracted and labeled in each form. Then, horizontal and vertical distances between each CC and its neighboring CCs are computed as:

$$x_{2i} - x_{1j} \quad i, j = 1,2,...,N \ , i \neq j \quad (1)$$

$$y_{2i} - y_{1j} \quad i, j = 1,2,...,N \ , i \neq j \quad (2)$$

where N is number of CCs and (x,y) is coordinate's of a CC bounding box. Figure 3 shows CC number 151 and its bounding box coordinate.

In the next step, labels of those CCs that their horizontal distance is smaller than a predefined threshold (X-threshold) are considered. If vertical distance between these blocks be also less than vertical threshold (Y-threshold), they are merged together to build the word block. After block merging, this process is repeated with new coordinates. Experimental results show that 10 iterations are enough to construct word block with several subwords.

X-threshold and Y-threshold are defined based on font size and lines distances. For example, X-threshold=15 pixels and Y-threshold=25 pixels are suitable for font size=14. Figure 4 shows the process of word block generation by concatenating CCs bounding boxes. Results of word blocking are shown in figure 5 for both handwritten and machine-printed words. It is obvious that handwritten word blocks overlap each other and have less regularity, compared with machine-printed word blocks. We will use such differences to separate handwritten and machine-printed words in the future.

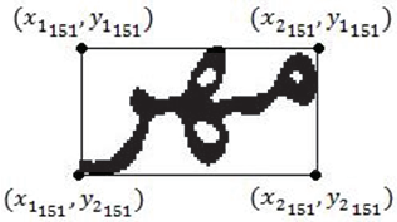


Figure 3. Bounding box of one CC

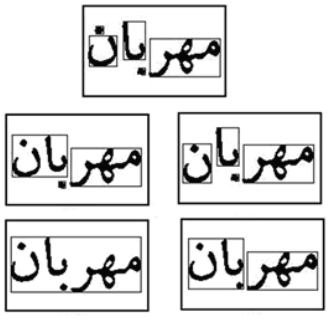


Figure 4. .Word block generation. First row: CCs bounding boxes. Second and third rows: merging of adjacent CCs bounding boxes.

C. Feature extraction

Three set of features are extracted from each word block to classify handwritten and machine-printed words. Since handwritten and machine-printed words have different visual appearances and physical structures, structural features are extracted to highlight these differences. Compared with handwritten documents, machine-printed texts often have simple stroke complexity. Therefore, crossing count histogram features are exploited as the second feature set. As the third feature set, we proposed baseline profile feature which is based on the properties of Farsi/Arabic languages. In the following sections, we present these features in detail.

1) Structural features

Previous efforts show that structural features are suitable to separate Latin handwritten and machine-printed words [1]. Likewise, two sets of structural features are extracted. The first set includes features related to the physical sizes of the blocks such as density of black pixels, width, height, aspect ratio, and area. The sizes of machine-printed words are more consistent than those of handwriting on the same form.

The second set of structural features is based on the connected components inside the block, such as the mean and variance of the width, height, aspect ratio, and area of the connected components. The sizes of connected components inside a machine printed word are more consistent, leading to smaller width and height variances. Unlike machine-printed words, for a handwritten word, the bounding boxes of the connected components tend to overlap with each other (figure 5). The overlapping area, normalized by the total area of the block is calculated as another feature (figure 6). We also use the variance of the vertical projection of each word block. Due to overlaps between handwritten blocks, projection profile has smoother valleys and peaks, resulting in smaller variance compared to machine-printed blocks (figure 7).

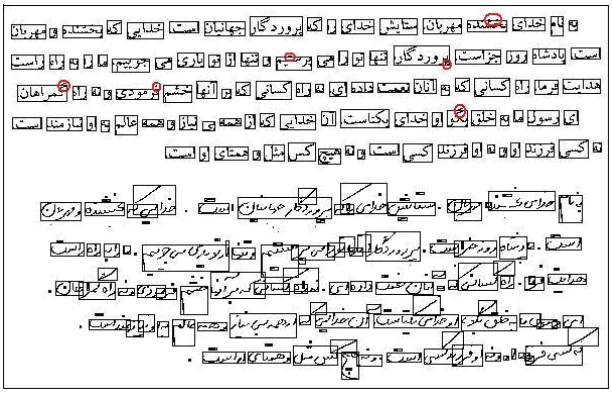


Figure 5. Results of word block for handwritten and machine-printed texts.

2) Crossing count histogram features

Like structural features, crossing count features were also used for Latin handwritten and machine-printed words discrimination [1]. Crossing count is the number of transitions from 0 to 1 along a hypothetical horizontal or vertical line over the word image. This feature can be used to measure stroke complexity. As shown in figure 8, the crossing counts of the first and second scan lines are 1 and 3, respectively. Horizontal and vertical crossing counts are defined as follows:

$$CCh(i) = \sum_{i} \overline{P}(i,j).P(i,j+1) \qquad (3)$$

$$CCv(j) = \sum_{j} \overline{P}(i,j).P(i+1,j) \qquad (4)$$

First, crossing counts in horizontal and vertical directions are extracted. To have scale-independent features, the obtained crossing count features are normalized.

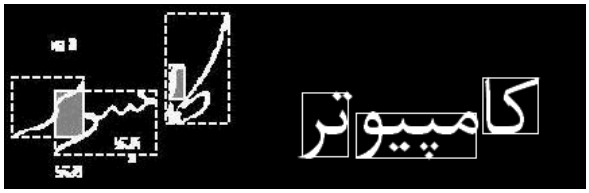


Figure 6. The overlap area of the connected components inside a word block for handwritten and machine-printed words.

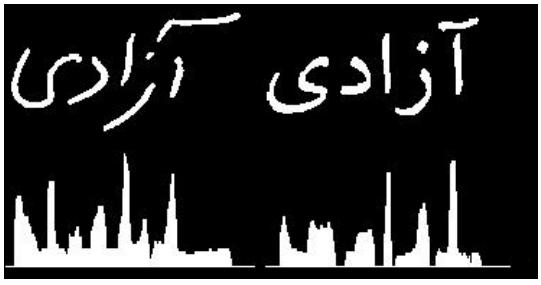


Figure 7. Valley and peak profiles of vertical projection for handwritten and machine-printed words.

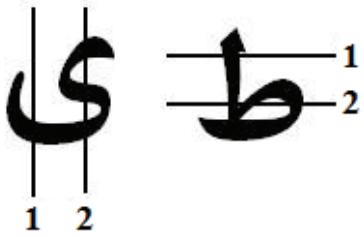


Figure 8. Crossing count features. The crossing counts of the first and second lines are 1 and 3, respectively.

Then histograms of horizontal and vertical crossing count features are calculated. For example, horizontal histogram of crossing count can be expressed as:

$$Ch_i = \sum_{k=1}^{w} G(k,u_i,\sigma)C'_k \qquad i = 1,2,3,4,5 \qquad (5)$$

Where w is the width of the block, C'_k denotes crossing count feature vector, and $G(k,u_i,\sigma)$ is a Gaussian-shaped function:

$$G(k,u_i,\sigma) = \exp\{-\frac{(k-u_i)^2}{2\sigma^2}\} \qquad (6)$$

Then, the histogram is divide into five bins with equal width and use five Gaussian-shaped weight windows to get the final features. σ is chosen so the weight on each bin border is 0.5. Five features are extracted in horizontal and vertical directions, leading to 10 features [1].

3) Baseline profile features

In this section we propose a new feature which is based on the properties of Farsi/Arabic languages. As explained in section 2.2 (word blocking), the utilized word blocking algorithm used horizontal and vertical thresholds extracted from machine-printed documents. Therefore, the word blocking algorithm results for handwritten words were not as accurate as machine-printed words (figure 5).

Difference between handwritten and machine-printed word blocks is used in the baseline profile features. Most pixels of a Farsi/Arabic word are placed on a hypothetical line, called baseline. In the machine-printed words, position of ascender (such as character ا) and descender (such as character ن) are determined by the baseline. However in handwritten case, words are not usually written on a single baseline. And position of ascender and descender varies according to the writer's style.

First, the baseline is estimated as the peak of horizontal histogram of the word image. Then, positions of highest and lowest scan lines are determined (line 1 and 2 in figure 9). Position of baseline in the word block, number of pixels on the baseline, and distances of highest and lowest scan lines from the baseline (d_1 and d_2 in figure 9) are considered as 4 features. Position of highest and lowest scan lines in a machine-printed word are more consistent, resulting smaller variances of d_1 and d_2.

Having baseline of each block, another 4 features can be extracted. First, we consider those pixels of the word image that lie on the baseline, called sub-baseline (figure 10). Then, number of obtained sub-baseline, mean and variance of them, and ratio of sub-baseline to their variances are taken as the last 4 features, leading to the final feature vector with 8 elements. In baseline profile feature, irregularity in the handwritten word blocks, and ascender and descender positions increase the ability of this feature to separate handwritten words from machine-printed words. Table 1 summarizes features used for Farsi/Arabic handwritten from machine-printed words discrimination.

TABLE I. FEATURES USED FOR HANDWRITTEN/MACHINE-PRINTED CLASSIFICATION.

Feature set	Size of feature vector
Structural	16
Crossing count	10
Base line	8
Total	34

D. Classification

As explained in the previous section, a total 34 features are extracted from each word block, consisting three different feature types.

The obtained feature vectors are fed into K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) classifiers to separate handwritten and machine-printed words. KNN classifier is faster for classification, and needs fewer training samples, while SVM classifier performs slightly better.

1) KNN classifier

As one would expect from the name, K-Nearest Neighbor classifies the input pattern by assigning it the label most frequently represented among the k nearest samples. In other words, a decision is made by voting the labels of the k nearest neighbors and taking a vote. In this paper, we used 1-NN and have assumed the Euclidean metric to determine neighboring patterns.

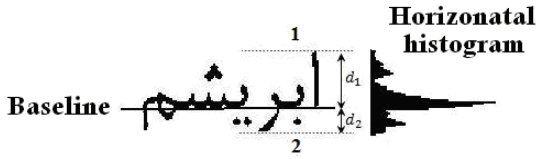


Figure 9. Baseline position and some features extracted from it.

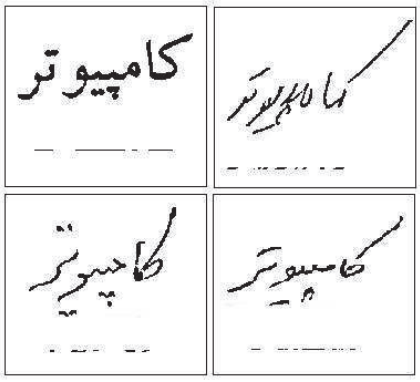


Figure 10. Sub-baselines in machine-printed and handwritten words.

2) SVM classifier

In Support Vector Machines, the input is mapped by a nonlinear function to a high dimensional space, and the optimal hyperplane found, the one that has the largest margin. The support vectors are those patterns that determine the margin; they are informally the hardest patterns to classify, and the most informative ones for designing the classifier.

Support vector machines outperform conventional classifiers, especially when the number of training data is small and the number of input variables is large. This is because the conventional classifiers do not have the mechanism to maximize the margins of class boundaries, resulting poor generalization ability is improved [9]. In this paper, we used a public SVM toolbox, called LibSVM, with RBF kernel [11].

III. EXPERIMENTAL RESULTS

In this paper we used 76 forms, designed for handwritten and machine-printed words discrimination (figure 2). They are written by different persons with different ages and educational backgrounds. Among them, 40 forms are used for training, and the other 36 forms for testing. As mentioned earlier (section 2.1), each form has an indicator line which separates handwritten and machine-printed parts; easing the task of training and testing classifiers. Using the word blocking algorithm results 32,006 blocks, containing 17,092 training blocks (including 9,095 handwritten blocks and 7,997 machine-printed blocks), and 14,914 testing blocks.

For performance evaluation, accuracy criterion is defined as follows:

$$Accuracy = \frac{\#\ of\ correctly\ classified\ blocks}{\#\ of\ blocks}$$

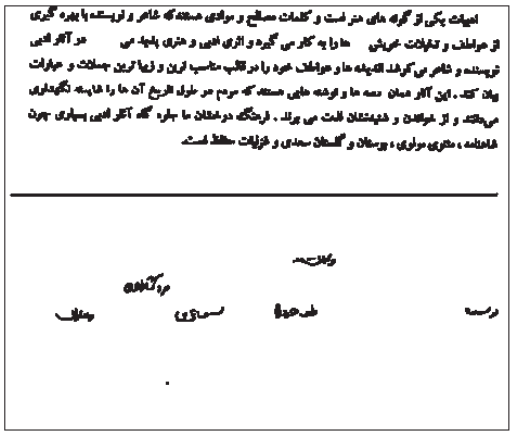
To maximize generalization ability of classifiers, we used cross validation technique. First, training set is divided into 10 random subsets (10-fold method). At each iteration, only one subset is used for testing, while the others have been used for training. This process is repeated ten times with different test set at each iteration. The average accuracy for all iterations is taken as the estimated accuracy for the current feature set (table 2). Afterwards, subsets with lowest error rates are used as the training set [10].

Table 3 compares classification time and classification accuracy of different features and classifiers over 14,914 blocks. Considering the same feature set, SVM classifier outperforms KNN classifier. However, KNN classifier is faster in both training and testing phases. Among three different features, structural features, with highest dimensionality, outperforms the others.

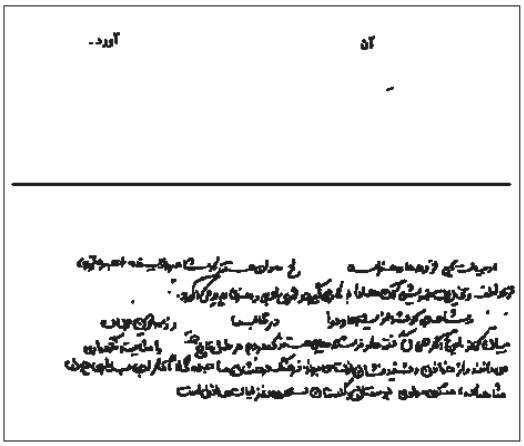
On the other hand, crossing count features has the lowest error variances. The low dimensional baseline profile feature is the most convenient feature to deal with. Another advantage of the proposed feature extraction method is its ability to separates handwritten and machine-printed words, especially when handwritten document is written carelessly. Figure 11 shows the results of Farsi/Arabic machine-printed and handwritten word separation using total feature set and SVM classifier.

TABLE II. MINIMUM ERROR RATES OF TRAINING SAMPLES WITH CROSS VALIDATION.

	Baseline profile features	**Crossing count features**	**Structural features**
K-NN Classifier	0.0819	0.0541	**0.0497**
SVM Classifier	0.0796	**0.0430**	0.0433



(a)



(b)

Figure 11. Word separation results. (a) machine-printed discrimination. (b) handwritten discrimination.

TABLE III. COMPARISON OF DIFFERENT FEATURES AND CLASSIFIERS FOR HANDWRITTEN/MACHINE-PRINTED WORDS CLASSIFICATION

	The k-NN classifier				The SVM classifier			
Feature	Correct	Accuracy	Variance	Testing Time (s)	Correct	Accuracy	Variance	Testing Time (s)
Structural	14094	94.5%	3.3%	56	14227	95.4%	4.3%	63
Crossing count	14019	94%	3.0%	41	14138	94.8%	3.2%	52
Baseline profile	13498	90.5%	10.4%	27	13597	91.2%	8.3%	38
Total features	**14398**	**96.5%**	**1.5%**	**71.4**	**14481**	**97.1%**	**0.9%**	**131.2**

IV. CONCLUSION

This paper presents a new method for Farsi/Arabic handwritten/machine-printed words classification. We proposed a new feature which uses irregularity of handwritten Farsi/Arabic words, appearing in the results of word blocking.

Combination of the baseline profile features with previously used structural and crossing count features results 34 features extracted from each block. Experimental results show that SVM classifier separates handwritten/machine-printed words with 97.1% accuracy.

REFERENCES

[1] Y.Zheng, H.Li, D.Doermann; "Machine-printed text and handwritten identification in noisy doxument images;" IEEE PAMI, 2004, Vol.26, No.3, pp.337-353.

[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, 1998, 31(9), pp.1275-1284.

[3] V.Pal and B.B.Chaudhuri, "Machine-printed and handwritten text lines identification", Pattern Recognition Letters, 22, pp.431-441, 2001.

[4] H.Ma and D.Doermann, "Gabor Filter Based Multiclass Classifier for Scanned Document Imagrs", Proc of 7th ICDAR, 2003, pp.968-972.

[5] Kavallieratou, E.; Stamatatos, S." Discrimination of machine-printed from handwritten text using simple structural characteristics" Proc of 17th ICPR, 2004, PP.437-440.

[6] Lorigo,L., Govindaraju, V., 2006. Offline Arabic handwriting recognition: a survey. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2006, Vol.28, No.5, , pp.712-724.

[7] M.Ziaratban, K.Faez, F.Bagheri," FHT: An Unconstraint Farsi Handwritten Text Database," Proc of 10th ICDAR, 2009, PP.281-285.

[8] Y.Zheng, H.Li, D.Doermann; "The Segmentation and Identification of Handwriting in Noisy Document Images;" Proc. IWDAS, 2002, pp. 95-105.

[9] V.Vapnik; "The Nature of Statistical Learning Theory;" Springer Verlag. New York: 1995.

[10] K.Fukunaga; "Introduction to Statistical Pattern Recognition;" second ed. Academic Press. New York: 1990.

[11] http://www.csie.ntu.edu.tw/~cjlin/libsvm/