Handwritten Character Recognition based on Structural Characteristics

E.Kavallieratou, N.Fakotakis and G.Kokkinakis
Wire Communications Lab.
University of Patras, 26500 Patras
ergina@wcl.ee.upatras.gr

Abstract

In this paper a handwritten character recognition algorithm based on structural characteristics, histograms and profiles, is presented. The well-known horizontal and vertical histograms are used, in combination with the newly introduced radial histogram, out-in radial and in-out radial profiles for representing 32x32 matrices of characters, as 280-dimension vectors.

The K-means algorithm is used for the classification of these vectors. Detailed experiments performed in NIST and GRUHD databases gave promising accuracy results that vary from 72.8% to 98.8% depending on the difficulty of the database and the character category.

1.Introduction

In this paper a handwritten character recognition approach is presented. The proposed technique has been developed as the last module of an integrated document analysis system, shown in fig. 1 [1-6].

In general, the character recognition procedure consists of two steps: (i) feature extraction where each character is represented as a feature vector and (ii) classification of the vectors into a number of classes. Govindan [7] distinguishes two categories of features: the structural and the statistical features, while Bunke [8] estimates that the structural approach is closer to the human way of recognition. In this paper, we propose a structural approach for feature extraction. Thus, a 280-dimension vector is extracted for each character, consisting of histograms and profiles. One new histogram and two new profiles are introduced. The k-means algorithm is, then, used for classification.

The proposed technique, described in section 2, is fast and simple, while the experimental results, illustrated in section 3, are quite promising. Finally, in section 4 our suggestions and plans for further improving are presented.

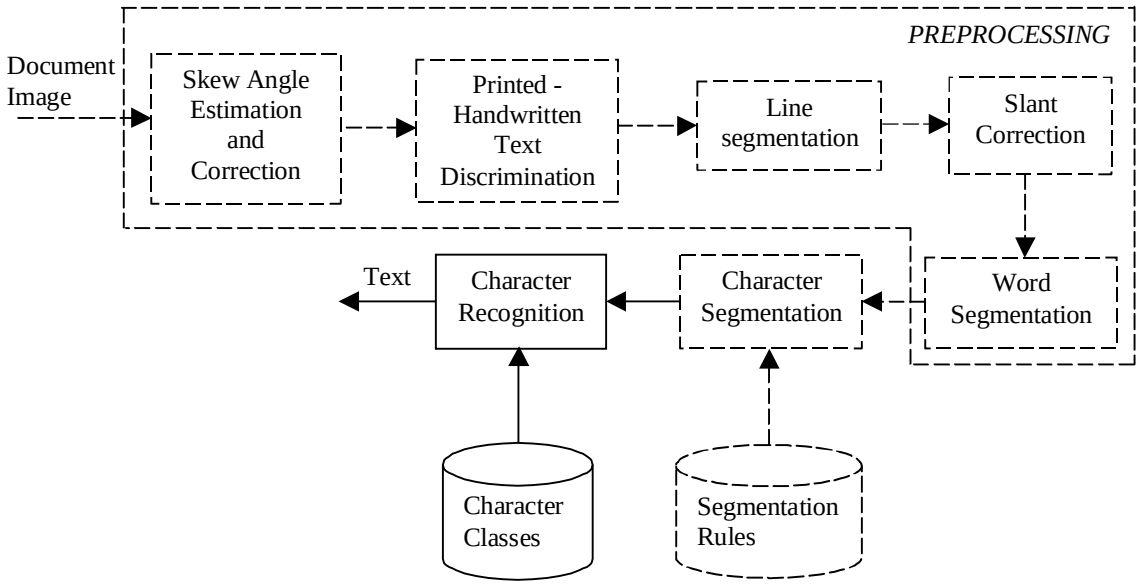
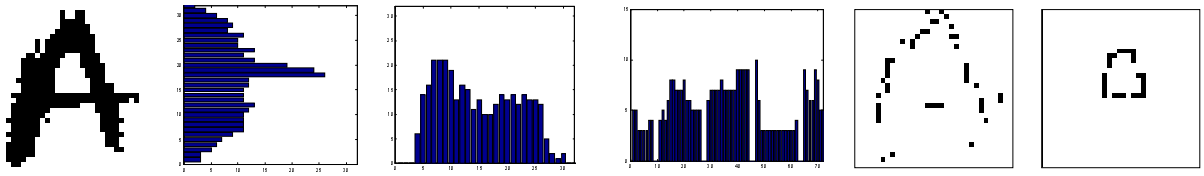


Figure 1. An integrated OCR system.



(a) (b) (c) (d) (e) (f)

Figure 2. a) Character matrix, b) Horizontal histogram, c) Vertical histogram, d) Radial histogram, e) Radial out-in profile and f) Radial in-out profile.

2. Algorithm description

As can been seen in fig. 1, after the preprocessing stage, isolated character images are produced which are used as input to the character recognition module. Each character is, then, represented as a 280-dimension vector, consisting of histograms and profiles. In more detail, each character is normalized in a 32x32 matrix. The horizontal histogram, the vertical histogram, the *radial* histogram, the *out-in* profile and the *in-out radial* profile are, then, calculated. The first two histograms are well-known and have been used extensively in optical character recognition research while the remaining features are used for first time. The formal definition of these features is follows.

Consider that the value of the element in the *m*-th row and *n*-th column of the character matrix is given by a function *f*:

$$f(m,n) = a_{mn}$$

where a_{mn} takes binary values (i.e., 0 for white pixels and 1 for black pixels). The horizontal histogram H_h of the character matrix is the sum of black pixels in each row (i.e., 32 features):

$$H_h(m) = \sum_n f(m,n)$$

Similarly, the vertical histogram H_v of the character matrix is the sum of black pixels in each column (i.e., 32 features):

$$H_v(n) = \sum_m f(m,n)$$

We define the value of the radial histogram H_r at an angle φ as the sum of black pixels on a rad that starts from the center of the character matrix (i.e., the element in the 16^{th} row and the 16^{th} column) and ends up at the border of the matrix, forming an angle φ with the horizontal axis. The radial histogram values are calculated with a step of 5 degrees (i.e., 72 features):

$$H_r(\phi) = \sum_{i=1}^{16} f(\|16 - i\sin\phi\|, \|16 + i\cos\phi\|),$$

$$\phi = 5 * k, k \in [0,72)$$

Additionally, we define the value of the *out-in radial profile* P_{oi} at an angle φ as the position of the first black pixel found on the rad that starts from the periphery and goes to the center of the character matrix forming an angle φ with the horizontal axis. The out-in radial profile values are calculated with a step of 5 degrees (i.e., 72 features):

$$P_{oi}(\phi) = \left\{ \begin{array}{l} I : \sum_{i=16}^{I-1} f(\|16 - i\sin\phi\|, \|16 + i\cos\phi\|) \equiv 0 \\ \& \; f(\|16 - I\sin\phi\|, \|16 + I\cos\phi\|) \equiv 1 \end{array} \right\},$$

$$\phi = 5 * k, k \in [0,72)$$

Similarly, we define the value of the *in-out radial profile* P_{io} at an angle φ as the position of the first black pixel on the rad that starts from the center and goes to the periphery of the character forming an angle φ with the horizontal axis. As above, the in-out radial profile values are calculated with a step of 5 degrees (i.e., 72 features):

$$P_{io}(\phi) = \left\{ \begin{array}{l} J : \sum_{i=0}^{J-1} f(\|16 - i\sin\phi\|, \|16 + i\cos\phi\|) \equiv 0 \\ \& \; f(\|16 - J\sin\phi\|, \|16 + J\cos\phi\|) \equiv 1 \end{array} \right\},$$

$$\phi = 5 * k, k \in [0,72)$$

An illustrated example is given in fig. 2, where the previously defined histograms and profiles (i.e., fig. 2b, 2c, 2d, and 2e) for a handwritten character (i.e., fig 2a) are shown.

Thus, a 280-dimension vector is extracted for each character image. A classification model is, then, produced by applying the k-means algorithm to the training data.

The preprocessing of the document images (see fig. 1) is likely to cause the undesirable segmentation of a handwritten character image into two character images. Taking this into account, during the classification of unseen cases, a feature vector is extracted for each

character image as well as for each pair of successive character images. These vectors are classified into the character class from which the Euclidean distance is minimized.

3. Experimental results

In order to evaluate our technique, we performed experiments using the NIST [9] database for English characters as well as the GRUHD [10] database for Greek characters. In both cases, the recognition system was trained using 2000 samples and 128 classes for each symbol and was tested on 500 samples for each symbol. The training and the test set were completely disjoint. Thus, the writers used in testing were completely different from the ones used in training.

In more detail, the recognition system was first trained based on the NIST database, for each one of the following categories separately: digits, uppercase characters and lowercase characters. Then, it was tested on unseen cases of the corresponding categories, taken from the same database. The accuracy rate in each case is shown in table 1. Since the output of the proposed character recognizer could be further improved by using lexicons, the recognition accuracy when the second and the third choices are taken into account are also given.

Next, the algorithm was trained and tested using the Modern Greek database of unconstrained writing (GRUHD). Since, the forms of GRUHD are very similar to those of the NIST database, it was possible to train our system separately for digits, uppercase and lowercase characters as well. The results on unseen cases of the corresponding categories are shown on table 2. In contrast to the NIST database, the GRUHD database contains unconstrained handwriting, so the accuracy results are somewhat lower in the latter case.

Table 1: The accuracy rates for the NIST.

	1st Choice	2nd Choice	3rd Choice
Digits	98.8%	99.91	100%
Uppercase characters	93.85%	96.54	98.86%
Lowercase characters	91.4%	94.50%	98.85%
Mixed characters	82.79	89.27%	96.85%

Table 2: Experimental results for the GRUHD.

	1st Choice	2nd Choice	3rd Choice
Digits	94%	97.42%	99,54%
Uppercase characters	86.03%	96.40%	98.96%
Lowercase characters	81%	90.36%	96.60%
Mixed characters	72.8%	80.04%	88.83%

In figures 3 and 4 the relation of the recognition accuracy with the training set size and the number of classes per symbol, respectively, is illustrated. In both cases the given results concern the characters of the NIST database.

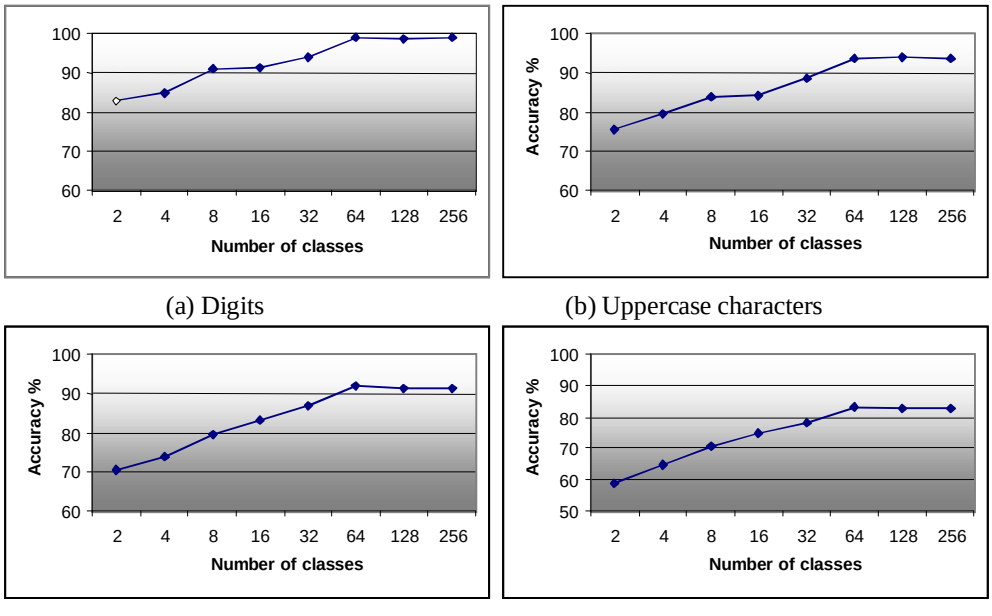
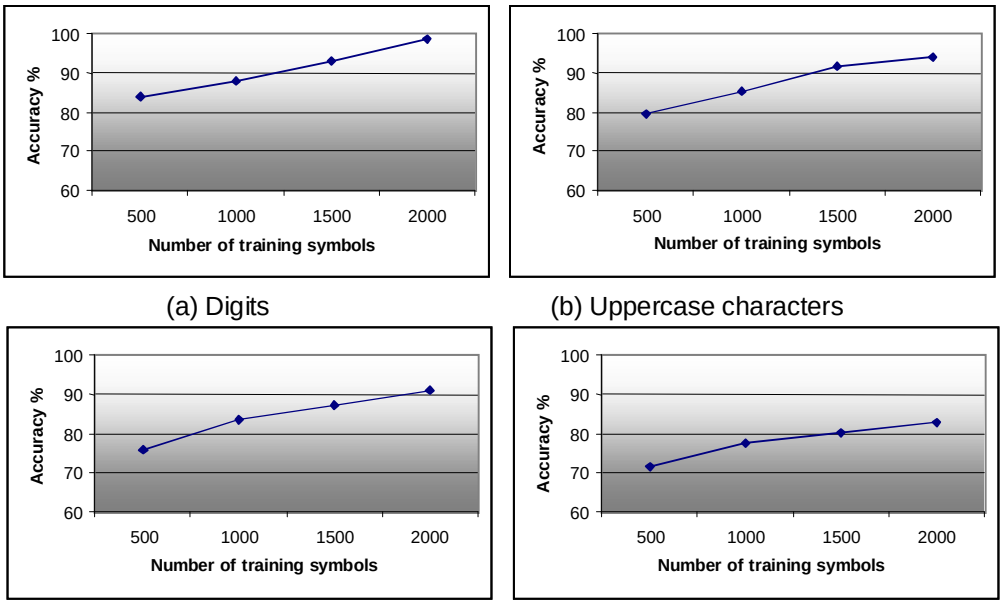


Figure 3. Recognition accuracy vs. number of classes per symbol for the NIST database.



(a) Digits (b) Uppercase characters

(c) Lowercase characters (d) Mixed characters

Figure 4. Recognition accuracy vs. training set size for the NIST database.

4. Conclusion

In this paper a technique for handwritten character recognition is presented. The proposed technique is focuses on the extraction of the features that best describe a handwritten character introducing one new histogram (i.e., radial) and two new profiles (i.e., in-out and out-in). These features together with the well-known horizontal and vertical histograms form a reliable representation of a handwritten character.

The described approach has been tested on two different databases with recognition accuracy varying from 72.8% to 98.8% depending on the difficulty of the database and the character category. The authors' conviction is that the lower results can be very much improved by the use of lexicons or grammar rules [11], which might be the subject of future work.

5. References

[1] E.Kavallieratou, N.Fakotakis and G.Kokkinakis, ``Skew Estimation using Cohen's class distributions", Pattern Recognition Letters 20, 1999, pp. 1305-1311.

[2] E.Kavallieratou, N.Fakotakis, and G.Kokkinakis, ``Slant Estimation Algorithm for OCR Systems", Pattern Recognition, v.34, n.12, 2001, pp.2515-2522.

[3] M.Maragoudakis, E.Kavallieratou, N.Fakotakis and G.Kokkinakis, ``How Conditional Independence Assumption Affects Handwritten Character Segmentaion", In Proc. ICDAR, 2001.

[4] E.Kavallieratou, D.C.Balcan, M.F.Popa, and N.Fakotakis, ``Handwritten Text Localization in Skewed Documents", In Proc. ICIP, 2001, pp.1102-1 105.

[5] E. Kavallieratou, E. Stamatatos, N. Fakotakis, and G. Kokkinakis, ``Handwritten Character Segmentation Using Transformation-Based Learning", In Proc. ICPR, 2000, pp.634-637.

[6] E.Kavallieratou, N.Fakotakis, G.Kokkinakis, ``New algorithms for skewing correction and slant removal on word-level", In Proc. ICECS, 1999, pp.1159-1162.

[7] V.K.Govindan, A.P.Shivaprasad, ``Character recognition -- A review", Pattern Recognition, v. 23, No 7, 1990, pp. 671-683.

[8] H.Bunke, A.Sanfeliu, Syntactic and structural Pattern Recognition, Theory and Applications, World Scientific, Singapore.

[9] R. Wilkinson, J. Geist, S. Janet, P. Grother, C. Burges, R. Creecy, B. Hammond, J. Hull, N. Larsen, T. Vogl, and C. Wilson, The first census optical character recognition systems conf. #NISTIR 4912. The U.S Bureau of Census and the National Institute of Standards and Technology. Gaithersburg, MD, 1992.

[10] E.Kavallieratou, N.Liolios, E.Koutsogeorgos, N.Fakotakis, G.Kokkinakis, ``The GRUHD database of Modern Greek Unconstrained Handwriting", In Proc. ICDAR ,2001.

[11] G.Frosini, B.Lazzerini, A.Maggiore, F.Marcelloni, ``Fuzzy classification based system for handwritten character recognition", In Proc. 2 nd Intern. Conf. On Knowledge-based Intelligent Electronic Systems (KES'98), pp. 61-65, 1998.