AMSE Review, AMSE Press, Vol. 17, N° 3, 1991, pp. 7-14
Received Aug. 16, 1990

A NEW METHOD FOR STATISTICAL CHARACTER RECOGNITION

N. Papamarkos, K. Tsirikolias and H. Spiliotis

Department of Electrical Engineering
Democritus University of Thrace
67100 Xanthi, Greece

Abstract

In the area of digital image processing an important subject is the optical character recognition (OCR). In the present paper, a fast OCR method is described for automatic reading typewritten characters. The proposed method is based on a statistical approach which consists of the use of the L_2 criterion in order to approximate the signature of the characters. The developed system is software-based using a microcomputer with 286 processor. The recognition rate is high in usually office applications.

1. Introduction

Optical character recognition is used to translate human-readable characters to machine-readable codes (for example ASCII). The main purpose of these OCR processes is to provide a fast input of documents to computers for storage and further processing. OCR has many scientific applications mainly in the area of text processing, office automation and computer aided design (CAD).

A number of approaches, which consider the OCR problem, have been proposed, the majority of which are based on statistical techniques, transforms and contour analysis [1-4]. All of these OCR methods have been carried out for typewritten or handwritten text.

This paper describes a new fast statistical OCR method for typewritten text. The first step of the new method is to determine

the signature of each typewritten letter. Then, the contour of the signature is approximated by using a fast approximation method. With this procedure we can obtain the features and the recognition coefficients of the characters. This technique may easily be implemented on a IBM PC (or compatible machinery), has a large recognition rate and low computing cost.

2. Image acquisition and encoding process

During this phase, the typewritten text is inserted to the computer by using of a high resolution binary scanner. By this way the text is transformed to a binary image which is divided into horizontal zones. Each zone contains the image form of a line of the text. The next step of the image processing is the encoding process by which the memory requirements for storing the image are substantial reduced. The encoding process is similar to the technique described in [5] but for binary images. Specifically, the mapping operation maps the sequence of image elements along a row $x_1,x_2,\ldots,x_n$ into a sequence of $g_1,g_2,\ldots,g_k$, where g_i denotes the run length of the same binary level. For example, the following sequence of pixels

$$\underset{\text{a}\quad\text{b}\quad\text{c}\quad\text{d}\quad\text{e}}{000\ldots000111\ldots100\ldots000111\ldots111000\ldots000} \qquad (1)$$

is stored as the abcde sequence of numbers, where

$$a+b+c+d+e = 512 \qquad (2)$$

In many cases, the application of the above procedure results to a reduction of the memory storage requirements greater than 90%.

Because an additional objective of the encoding process is for the text line division, each image must satisfy the following constraints:

1. It is necessary that between two text lines there is a line with zero grey level.
2. The characters in each line must not tangent each others

3. The first column of image must has only pixels of zero level. The horizontal image length is 512 pixels

3. Signature extraction

After the encoding process the text image is stored as a data file for the following recognition processing:

- Text lines of the image are separated by using the data file information. Each of this lines represents the input of the next recognition process.

- Each character of a line is selected from the left to the right direction, scaled and normalized to an 64X64 image. The normalization procedure is independent of the size of the characters.

The next step for the features extraction is the transition of the two-dimensional character representation to a corresponding one-dimensional which is known as the signature of each character. A simple method that permits to generate signatures is to plot the distance from the center-point to each pixel of the character as a function of the pixels serial numbers. The next step is the approximation of the signature by using m times an algorithm which is based on the L_2 criterion. At every iteration of the algorithm the signature is approximated by a different number of coefficients such as the final objective errors to correspond to the characters recognition coefficients. The above algorithm is fast and can be easily implemented. The number of m iterations depends on the specific application and of the noise of the binary image.

4. Problem formulation

According to the above, the proposed OCR method consists of the following steps:

Step 1. The coordinates of the center-point of each scaled character are represented as (X,Y) where X and Y are equal to:

$$X = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad \text{and} \qquad Y = \frac{1}{N}\sum_{i=1}^{N} Y_i \tag{3}$$

N is the total number of pixels and X_i, Y_i are the coordinates of each pixel.

Step 2. In this step we determine the distances y_i between the (X,Y) point and the pixels of the character. For every pixel i, these distances are given by the relation:

$$y_i = \sqrt{(X-X_i)^2 + (Y-Y_i)^2} \tag{4}$$

It is clear that the values of xi and y_i represents the 1-D curve, i.e. the signature of each character.

Step 3. This is the approximation step in which the signature of each character is approximated m times. The value of m is depended of the number of the necessary recognition coefficients.

The x-axis of the one-dimensional curve of each signature represents the serial number of the pixels of the character, i.e. $x_i=i$ for i=1,2,...,N. Also, in the y-axis let that y_i represent the distance of pixel i from the center-point. Our scope is to approximate the signature of the character by an appropriate polynomial of the following form:

$$q(x_i) = a_1 + a_2x_i + a_3x_i^2 +...+ a_Mx_i^{M-1} \tag{5}$$

where a_i, i=1,2,...,M are the unknown coefficients of the polynomial and M<N. By applying the L_2 criterion the approximation problem leads to minimize the quantity:

$$e = \sum_{i=1}^{N} \left[q(x_i) - y_i\right]^2 \tag{6}$$

It is well known that the minimization of e leads to the solution of a set of M linear equations [6]. It is clear that most sophisticated algorithms can be applied in this step, as for example the rational approximation algorithm [7-8].

Step 4. If e_k, k=1,2,...,m is the minimum value of e in the k iteration, then the minimization of e m times will create a vector E equal to:

$$E = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{m+1} \end{bmatrix} \tag{7}$$

where the e_{m+1} recognition coefficient is given by the relation

$$e_{m+1} = \begin{bmatrix} e_1 & e_2 & \cdots & e_m \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix} \tag{8}$$

The value of each element e_k represents a character recognition coefficient.

Step 5. In this step the character recognition coefficients e_k, $k=1,2,\ldots,m+1$ are used in order to give us a decision about the character. For this reason we use a look-up table in which we have store the prototype values of the recognition coefficients p_k for each character. The Euclidean distance between the recognition coefficients is given by

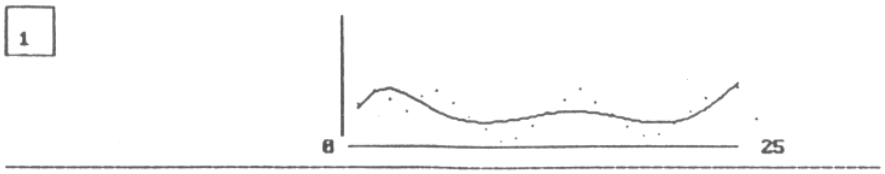
$$D = \sum_{k=1}^{m+1} \left[p_k - e_k \right]^2 \tag{9}$$

Next, the value of D is compared with appropriate threshold values and give us the decision about the character. It is noted that the threshold values for D are determined initially by using an iterating test procedure.

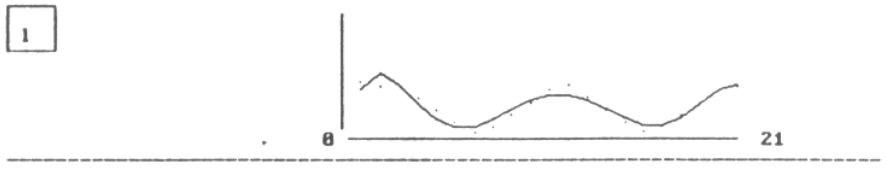
5. Examples

Example 1

In this example the character recognition procedure is applied to the letters 1 and l. These letters have common figure characteristics which results in similar signatures. The approximation of these signatures by polynomials of six order, gives us the results which are showed in Figure 1. It is noted that the two polynomials have similar schemes but the difference of the approximation errors (recognition coefficients) is significant.




```
error= 28.58624196341314

COF
-.3495789684358123
 4.490944724217309
-1.424746972827219
 .1848291987553461
-1.139727186923617E-002
 3.357830648441147E-004
-3.764173415710356E-006
```



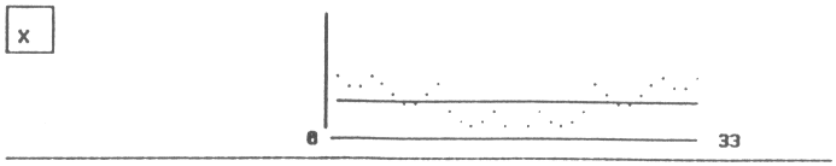

```
error= 4.488493962253851

COF
-1.842746974731697
 9.089184121948439
-3.957873495723538
 .6833113042959355
-5.555188722631092E-002
 2.129939326799662E-003
-3.184556322894276E-005
```

Figure 1. Approximation of the signatures of the characters 1 and 1 by six order polynomial

Example 2

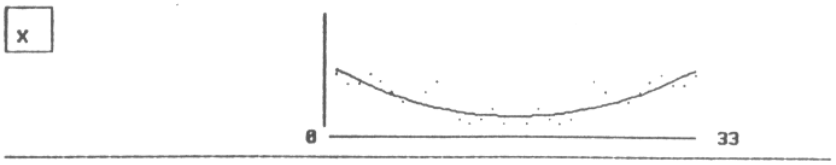
This example illustrates the application of the recognition algorithm to the X character. The algorithm is applied for m=4 and N=33. The complete approximation results are given in Figure 3 which gives the following values for the recognition coefficient of the vector E:

$$E = \begin{bmatrix} 69.82190 \\ 21.86115 \\ 16.48445 \\ 15.52994 \\ 5865.93154 \end{bmatrix} \qquad (10)$$



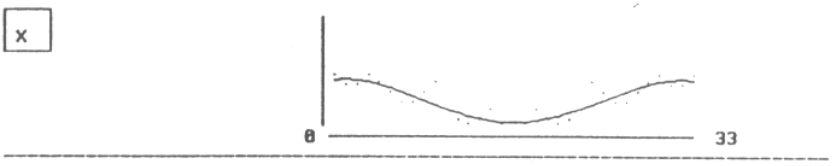
error= 69.82195616788385

COF
3.085943828145687



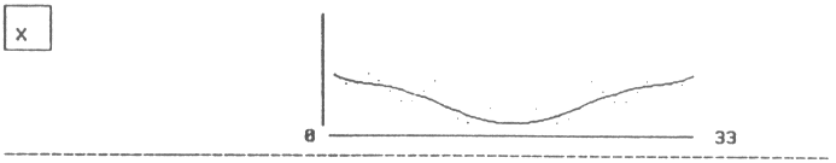
error= 21.861152519871 74

COF
5.958437885284186
-.5861416875894572
1.488652822886639E-002



error= 16.48445827988288

COF
4.317346969652444
.3453885767962557
-9.393762344712263E-002
4.928198385288563E-003
-7.247358448824389E-005



error= 15.529935240368 97

COF
5.488881988371521
-.7598821284797 56
.2827183226991232
-2.859817590828888E-002
1.743226384229281E-003
-4.668229319758242E-005
4.576695411538683E-007

Figure 2. Polynomial approximations of the character X

6. Conclusions

This paper describes a new fast statistical method for optical character recognition. The simplicity and the effectiveness of the proposed method make it suitable for practical typewritten OCR systems. After the processes of acquisition filtering and scaling, the new method in its first step, produces an one-dimensional curve which is the signature of the typewritten characters. Next, the signature is approximated iteratively by a fast algorithm which is based on the L_2 criterion. The product of this procedure is a vector, the elements of which are considered as recognition coefficients. The proposed technique can be used easily as a part of a more complex syntactic recognition method and can be optimized by using rational functions approximation fast algorithms.

7. References

[1] G.L. Cash and M. Hatamian, "Optical character recognition by the method of Moments," *Computer Vision Graphics and Image Processing*, Vol. 39, pp. 291-310, 1987.

[2] S. Kahan, T. Pavlidis and H.S. Baird, "On the recognition of printed characters of any font and size," *IEEE trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-9, No. 2, March 1987.

[3] G. Nagy, *Optical character recognition--Theory and practice,* in Handbook of Statistics, Vol. 2., P. R. Krishnaiah and L.N. Kanal, Eds. Amsterdam, The Netherlands: North-Holland, pp. 621-649, 1982.

[4] F.W.M. Stentiford, "Automatic feature design for Optical Character recognition using an evolutionary search procedure," *IEEE trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-7, pp. 349-355, 1985.

[5] R.C. Gonzalez and P. Wintz, *Digital Image Processing*, Second Edition, Addison-Wsley Pub. Company, May 1987.

[6] W.H. Press, B.P Flannery, S.A. Teukolsky and W.T. Vetterling, *Numerical Recipes-The Art of Scientific Computing*, Cambridge University Press, 1986.

[7] N. Papamarkos, G. Vachtsevanos and B. Mertzios, "On the optimum approximation of real rational functions via linear programming", *Applied Mathematics and Computation*, Vol. 26, pp. 267-287, 1988.

[8] N. Papamarkos, "A program for the optimum approximation of real rational functions via linear programming", *Advances in Engineering Software*, Vol. 11, No. 1, January 1989.