# Character Recognition Using Canonical Invariants

Sema Doguscu and Mustafa Unel

Department of Computer Engineering, Gebze Instute of Technology
Cayirova Campus
41400 Gebze/Kocaeli Turkey
{doguscu, munel}bilmuh@gyte.edu.tr

**Abstract.** This paper presents a new insight into character recognition problem. Implicit polynomial (IP) curves have been used for modelling characters. A unique decomposition theorem is employed to decompose these curves into simple line primitives. For the comparison of the characters, canonical invariants have been computed using so called "related points" of the curves, which are the real intersections of the lines. Experimental results are presented to asses discrimination power of proposed invariants and their robustness under data perturbations. The method has also been compared with fourier descriptors.

## 1 Introduction

Automatic recognition of characters is an important problem in pattern analysis, and it has been the subject of research for many years. This paper presents a new insight into character recognition problem using IP or so called algebraic curves. The problem is to assign a digitized character into its symbolic class. In this work, IP curves have been used to model characters. Implicit polynomials are one of the most effective representations for complex free-form object boundaries and have certain advantages over other representations [4,5,6,7,8,10]. Character recognition follows three major steps in our approach. These are [1]: Preprocessing; Representation; Recognition and Classification of Characters.

In the preprocessing part, analog documents are converted into digital form and then thresholded. The connected component analysis [2] is performed on digitized image and each character is extracted from the text line. Then boundaries of segmented characters are obtained by eight-neighbor method [3]. In the representation part, characters are modelled using IP curves which are fitted to the boundaries of the characters by a fitting procedure [4]. A unique decomposition theorem [6] is then used to decompose algebraic curves into lines. Line factor intersections are related- points which map to one another under affine transformations. These related-points are used to construct canonical invariants [7], which will then be used in recognition and classification part. In recognition and classification part, characters are recognized by comparing their canonical invariants. To compare invariant vectors, a similarity ratio is employed.

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3211, pp. 432–439, 2004.

## 2 Preprocessing

The raw character data are subjected to a number of preliminary processing steps. These preprocessing algorithms smooth the character images, segment the characters from each other and from the background, remove the noise and calculate the boundaries of characters. The scanned input character images are gray-scaled images. These images should be converted into binary images by *thresholding* [3]. See Fig. 1.b. For representation, characters should be isolated from the document and each other. *Segmentation* is division or separation of the image into regions of similar attribute. Connected component segmentation method [2] is used in this work. Fig. 1.c depicts some segmented characters of the document shown in Fig. 1.a.

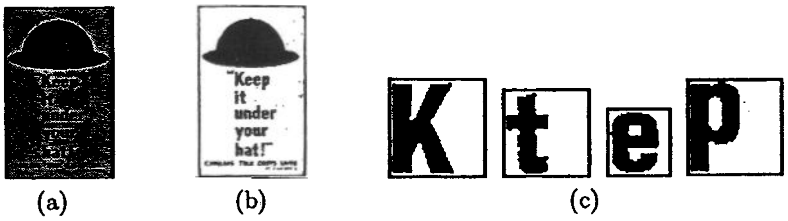

**Fig. 1.** (a)Original image (b)Binarized Image (c)Segmented characters

Since each character can be represented by a closed curve contour of line segments, tracing the boundary of the character can yield useful information to distinguish characters from one another [9]. *Contour detection* algorithm [3] extracts the information of the boundary of a segmented character and presents it in a more compact form. See the boundaries of some characters in Fig. 2.

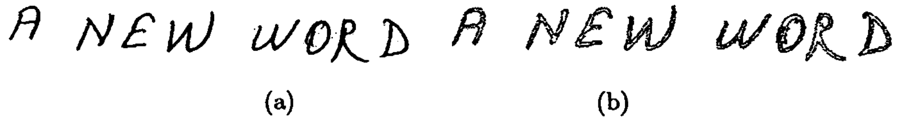

**Fig. 2.** (a)Original Image (b)Character Contours

## 3 Representation

### 3.1 Algebraic Curve Fitting and Implicit Polynomials

Image representation plays one of the most important roles in a recognition system. In order to avoid extra complexity and to increase the accuracy of the algorithms, representation is required. In this work, algebraic curves have been used for image representation. To get the best fitting polynomial, we have used a fitting algorithm detailed in [4]. This algorithm is linear, computationally fast, Euclidean invariant and robust. Examples of algebraic curve fits of $6^{th}$ degree to data sets are shown in Fig. 3. Our experiments have shown that virtually all of the characters can be fit well by sixth degree IP curves.

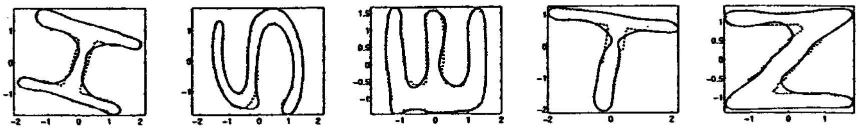

**Fig. 3.** Algebraic curve fitted characters. Solid curves represent algebraic curves fitted to the characters.

IP curves and surfaces are mathematical models for the representation of 2D curves and 3D surfaces. Algebraic curves are defined implicitly by equations of the form $f(x,y)=0$, where $f(x,y)$ is a polynomial in the variables $x, y$, i.e.

$$f(x,y)=\sum_{0\leq i+j\leq n} a_{ij}x^i y^j$$

Alternatively, the intersection of an explicit surface $z=f(x,y)$ with the $z=0$ plane yields an algebraic curve if $f(x,y)$ is a polynomial [6]. The analysis and representation of algebraic curves can be simplified through *a unique decomposition theorem.* Decomposition theorem provides a new expression for the curve as a unique sum of the products of (possibly complex) lines.

**Theorem 1.** *[6] A non-degenerate (monic) $f_n(x,y)$ can be uniquely expressed as a finite sum of real and complex line products or real conic-line products, namely*

$$f_n(x,y)=\Pi_n(x,y)+\gamma_{n-2}[\Pi_{n-2}(x,y)+\gamma_{n-4}[\Pi_{n-4}(x,y)+\ldots]] \tag{1}$$

*where each $\Pi_r$ is a product of r real and/or complex lines,i.e.*

$$\Pi_r(x,y)=\Pi_{i=1}^{r}(x+l_{ri}y+k_{ri}) \text{ with } \Pi_0(x,y)=1$$

### 3.2 Affine Equivalence and Related Points

Any two curves defined by a monic $f_n(x,y)=0$ and a monic $\bar{f}_n(\bar{x},\bar{y})=0$ will be affine equivalent if for some scalar $s_n$,

$$f_n(x,y)=0\overset{A}{\mapsto} f_n(m_1\bar{x}+m_2\bar{y}+p_x, m_3\bar{x}+m_4\bar{y}+p_y)\overset{def}{=} s_n\bar{f}_n(\bar{x},\bar{y})=0 \tag{2}$$

Two corresponding related-points of the affine equivalent curves defined by $f_n(x,y)=0$ and $\bar{f}_n(\bar{x},\bar{y})=0$, such as $\{x_i,y_i\}$ and $\{\bar{x}_i,\bar{y}_i\}$ will be defined by the condition that

$$\begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} m_1 & m_2 & p_x \\ m_3 & m_4 & p_y \\ 0 & 0 & 1 \end{pmatrix}}_{A} \begin{pmatrix} \bar{x}_i \\ \bar{y}_i \\ 1 \end{pmatrix} \longrightarrow \{x_i,y_i\} \overset{A}{\mapsto} \{\bar{x}_i,\bar{y}_i\} \tag{3}$$

Any two corresponding related-points will satisfy the relation

$$f_n(x_i,y_i) = s_n \bar{f}_n(\bar{x}_i,\bar{y}_i) \tag{4}$$

In the case of affine transformations, bitangent points, inflection points, centroids and line factor intersections all represent related points which can be determined from knowledge of the curves [6]. Line factor intersections have been used as related-points in this work. To establish the *correct correspondence* between the points in two sets of $k$ corresponding real, distinct related-points, such as $\{x_i,y_i\}$ and $\{\bar{x}_i,\bar{y}_i\}$, we next note that if $f_n(x_i,y_i)=z_i$ and $\bar{f}_n(\bar{x}_i,\bar{y}_i)=\bar{z}_i$, then $z_i = s_n\bar{z}_i$ and

$$s_n = \frac{z_i}{\bar{z}_i} = \frac{\sum_{i=1}^{k} z_i}{\sum_{i=1}^{k} \bar{z}_i} \tag{5}$$

Therefore, we will always order the related points so that $z_1 < z_2 < ... < z_k$, and

$$\bar{z}_1 < \bar{z}_2 < ... < \bar{z}_k \quad if \quad s_n > 0$$

$$\bar{z}_1 > \bar{z}_2 > ... > \bar{z}_k \quad if \quad s_n < 0$$

## 4 Recognition

To distinguish characters from one another, a set of features should be extracted for each class and these features should be invariant to characteristic differences within the class. In this work *canonical invariants* have been used for recognition. Let $f_n(x,y)=0$ and $\bar{f}_n(\bar{x},\bar{y})$ be affine equivalent IP curves. Any three related-points of $f_n(x,y)=0$ to any three corresponding related-points of $\bar{f}_n(\bar{x},\bar{y})=0$ will define the affine transformation matrix A via the relation (3).

Any three such related-points of $f_n(x,y)=0$ will define a canonical transformation matrix [7]

$$A_c = \begin{pmatrix} x_1-x_3 & x_2-x_3 & x_3 \\ y_1-y_3 & y_2-y_3 & y_3 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{pmatrix} \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix}}_{E}, \tag{6}$$

and a monic canonical curve $f_n^c(x,y)=0$ of $f_n(x,y)=0$ defined by the relation

$$f_n(x,y)=0 \overset{A_c}{\mapsto} s_c\, f_n^c(x,y)=0 \tag{7}$$

Three corresponding related points of $\bar{f}_n(\bar{x},\bar{y})=0$ will define a corresponding canonical transformation matrix $\bar{A}_c=\bar{T}E$, and a corresponding monic canonical curve $\bar{f}_n^c(\bar{x},\bar{y})=0$ of $\bar{f}_n(\bar{x},\bar{y})=0$ defined by the relation

$$\bar{f}_n(\bar{x},\bar{y})=0 \overset{\bar{A}_c}{\mapsto} \bar{s}_c\, \bar{f}_n^c(\bar{x},\bar{y})=0 \tag{8}$$

for some scalar $\bar{s}_c$

We will call the coefficient vectors of canonical curves *canonical invariants*. Our strategy will be to associate a canonical curve with each character and compare characters based on their canonical curves. In practice, the coefficients of the canonical curves will not be the same for the affine equivalent characters because of noise. Therefore we have to introduce some measure of "closeness" of canonical invariant vectors. Comparison of two characters is realized by comparing the similarity of the canonical invariants. Charàcters have been compared with each other under a similarity ratio. The similarity ratio employed in this work is

$$Similarity = r = \frac{C_1^T C_2}{\|C_1\|\|C_2\|} \tag{9}$$

Here $C_1$ and $C_2$ are the canonical invariants of the curves. If two vectors are close to each other, similarity value gets closer to 1, otherwise similarity value gets closer to -1. Characters with the highest similarity will be considered to be equivalent, and therefore to be the same.

$$-1 \leq r \leq 1 \tag{10}$$

## 5 Experimental Results

We now present some experimental results which illustrate our procedures. Characters have first been thresholded, segmented and their boundaries have been extracted. Then IP curves have been fit to data sets. After obtaining three related-points from IP curves, the (monic) canonical curves $f_6^c(x,y)=0$ have been determined. Using canonical invariant vectors, similarity ratio between the characters has been computed.

Recognition is performed by comparing the input characters with various model characters in the database using the computed similarity ratios. Characters can be classified into 3 groups by the number of their contours. Each character in the first group has one contour (See fig. 4a). The ones in the second group has two contours (See fig. 4b). Those in the third group has three contours as shown in fig. 4c. Several characters have been tested and their similarities to the model characters have been computed . The character model which has the

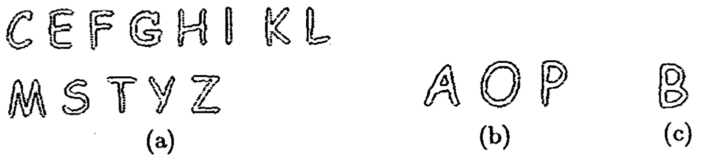

**Fig. 4.** (a) First group (b) Second group (c) Third group

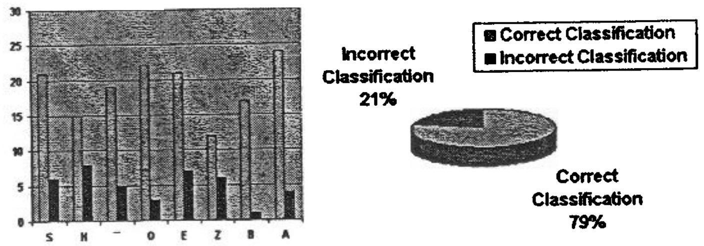

**Fig. 5.** Recognition rates using implicit polynomials

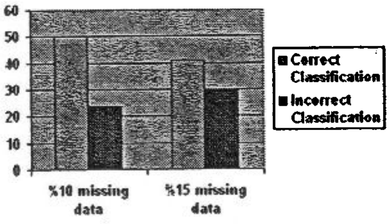

**Fig. 6.** Correct and incorrect classification rate under 10% and 15% missing data using implicit polynomials. Canonical invariants have yielded 68% recognition rate under10% missing data and 58% recognition rate under 15% missing data.

largest similarity ratio has been declared as the input character. We have used 191 character data and the recognition rate was 79%. See Fig. 5.

A character recognition system usually doesn't have all the boundary information of the characters. Only partial information might be available. To test the robustness and the discrimination power of our canonical invariants with respect to missing data, character data points were chopped at different boundary locations. The similarity ratios based on the canonical invariants of characters under 10% and 15% missing data are computed and shown in Fig. 6. Canonical invariants have yielded 68% recognition rate under10% missing data, and 58% recognition rate under 15% missing data.

We have also compared our method with fourier descriptors using the same characters, same models and the same conditions. Characters have been thresholded, segmented and their boundaries have been extracted. Then fourier descriptors have been computed. From these descriptors, similarity ratios have been computed. Recognition rate has been found to be 69%.

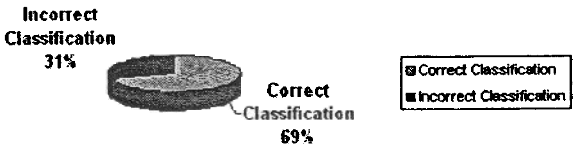

**Fig. 7.** Recognition rate for fourier descriptors using the same characters, the same models and the same conditions.

To test the robustness and the discrimination power of fourier descriptors with respect to missing data, character data points were chopped at different boundary locations. The similarity ratios of characters under 10% and 15% missing data have been computed and shown in Fig. 8.

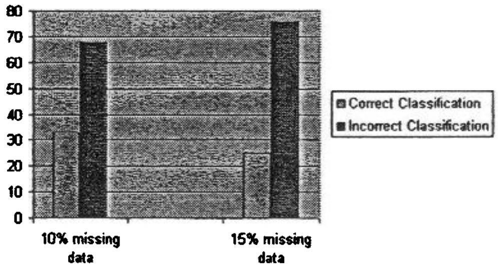

**Fig. 8.** Correct and incorrect classification rate under 10% and 15% missing data using fourier descriptors. Fourier descriptor based invariants have yielded 33% recognition rate under10% missing data and 25% recognition rate under 15% missing data.

## 6 Conclusion

We have now outlined a new method for character recognition problem. Algebraic curves are used for modelling characters. Most of the characters can be represented by $6^{th}$ degree algebraic curves. Since the quality of the fitting algorithm has substantial impact on the recognition performance, a stable and repeatable curve fitting method has been used. Decomposition theorem is employed to decompose these curves into lines. Line factor intersections have been used as related-points. By using related-points, canonical invariants have been computed.

Experiments have been conducted to compare characters based on the similarity between canonical invariant vectors. Robustness and the discrimination capabilities of canonical invariants have been tested on different characters. Experiments have shown that canonical invariants are stable with respect to modest amount of missing data. We have also compared our method with fourier descriptor using the same characters, the same models and the same conditions. Experimental results are promising, and much work must be done to fully exploit advantages of using IP curves as a representation in character recognition problems.

**Acknowledgment.** This research was supported from GYTE research grant BAP #2003$A$23

## References

1. N. Arica & F. Yarman-Vural, An Overview of Character Recognition Focused on Off-Line Handwritting, IEEE Transactions on Systems,Man and Cybernetics-Part C:Applications and Reviews,Vol.31,No.2, May 2001.
2. H. Kuo & J. Wang, A New Method for the Segmentation of Mixed Handprinted Chinese/English Characters, Proceedings of the Second International Conference on Document Analysis and Recognition, pages 810-813, October 1993.
3. C. Jeong & D. Jeong, Handwritten Digit Recogntion Using Fourier Descriptors and Contour Information, IEEE TENCON, vol.6, No.99, 1999.
4. T.Tasdizen & J.P. Tarel & D.B. Cooper, Improving the Stability of Algebraic Curves for Applications, IEEE Transactions in Image Processing, vol.9, No.3, March 2000.
5. M. Unel & W. A. Wolovich, A new representation for quartic curves and complete sets of geometric invariants, International Journal of Pattern Recognition and Artificial Intelligence, December 1999.
6. M. Unel & W. A. Wolovich, On the Construction of Complete Sets of Geometric Invariants for Algebraic Curves, Advances in Applied Mathematics, Vol. 24, No. 1, pp. 65-187, January 2000.
7. W. A. Wolovich & M. Unel, The Determination of Implicit Polynomial Canonical Curves, IEEE Transactions on Pattern Analysis and Machine Intelligence, October 1998.
8. M.Blane, Z.Lei et al., The 3L algorithm for Fitting Implicit Polynomial Curves and Surfaces to Data, IEEE Transaction on Pattern Analysis and Machine Intelligence, Bol.22, No.3, March 2000.
9. Y. Chung & M. Wong, Handwritten Character Recognition by Fourier Descriptors and Neural Network, IEEE TENCON, Speech and Image Technologies for Computing and Telecommunications, 1997.
10. D. Keren & D. Cooper, Describing Complicated Objects by Implicit Polynomials, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.16, No.1, 1994.