Multi-font English Character Recognition based on Modified Invariant Moments

[1]B. V. Dhandra, [1]V. S. Malemath*, [1]Mallikarjun H., [1]Ravindra Hegadi
[1] Post-Graduate Department of Studies and Research in Computer Science,
Gulbarga University, Gulbarga-585 106, India.
E-Mail: dhandra_b_v@yahoo.co.in, veeru_sm@yahoo.com*

Abstract: This paper describes an approach based on modified invariant moments for recognition of multi-font English characters. The proposed method is independent of size and translation variations and showed better results under noisy conditions. The work treats isolated English characters which are normalized to a size of 33 × 33 pixels and the image is thinned. As a pre-classification step end points and Euler numbers have been estimated from this thinned image of the character. For size and translation invariance the modified invariant moments suggested by Palaniappan have been evaluated. The system is trained for 7 different font-styles with 364 images. A decision tree based minimum distance nearest neighbor classifier has been adopted for classification. The system is tested for these seven fonts with various sizes of the characters between 8 to 72. The total of 7,280 character images are tested with this system and the success rate is found to be 99.65%. The method shows encouraging results on multi-font/sized character images.

Keywords: Character recognition, modified invariant moments, multi-font, end point, Euler number.

1. Introduction

Optical character recognition (OCR) is one of the thrust areas of research in pattern recognition due to its potential applications. Today, it is possible to have reasonably good OCR package in the market, but recognition rate often drop drastically whenever the font style and or the size of the character changes. Application of multi-font and variable size character recognition can be found in office automation, computer aided design and many other related areas. This has motivated us to develop the methods which are independent of font/size constraints, to achieve the desired level of accuracy and to have faster rate of recognition.

The general character recognition system often consist of 4 stages viz. pre-processing, normalization, feature extraction and classification. The Fig. 1 shows flow of the steps involved in a general recognition system.

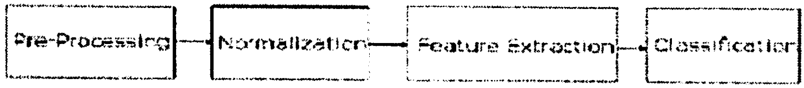

Fig. 1 Block diagram of general recognition system

After acquiring the images from the scanners the first stage will be a preprocessing step in which the noise removal, skew/slant correction is often performed as the images captured from the scanners often have noise and or skewed. The second step is the normalization as the input character images are often in different font sizes. Normalization to a particular size makes it easy for the feature extractor. After normalization these character images are subjected to thinning or skeletonising.

Selection of features extraction method is probably the single most important factor in achieving high recognition performance. Feature extraction plays a very vital role especially in multi-style environment as it has to minimize the within class pattern variability while enhancing the between class pattern variability. In the past researchers made use of different type of tools for feature extraction. The literature [3, 7, 11, 13, 15, 19] reveals dynamic programming, Hidden markov models (HMM), template matching, moment functions, Hough transform, Fourier descriptors, hybrid, statistical/syntactical features etc. Finally, in the classification stage which assigns the input image to a class of best match with the known class. The different classifiers used are nearest neighbor (NN), k-NN, Bayes classifier, Fuzzy, Neural networks, Support vector machines (SVM) etc.

The prominent among the feature extraction techniques are moments of the image. Moment functions have a very large set of potential applications in image analysis, viz. invariant pattern recognition, object/script recognition, pose estimation etc. A set of moments generally represents global characteristics of the image shape and provides information about the different types of geometric features in the image. One of the significant papers based on moment invariants was proposed by Hu [9] in 1962. This paper appears in all the citations of moment related papers till the date. In this paper the central moments and mass normalization technique are used to obtain invariance to translation and scaling. Hu also proposed two different algorithms to obtain rotation invariance viz. the combinations of regular moments using algebraic invariants which are known as the absolute moment invariants and the principal axis method. These together are known as Hu's moments. Hu's absolute moments have been employed in several character recognition works. Various types of moments have been used to recognize image patterns in a number of applications. Moments considered include regular moments [10], Geometric [6, 20], Legendre moments [4], Zernike/pseudo-Zernike moments [12, 18, 19], affine [5], radial and rotational moments [2, 17], and complex moments [1]. Recently Palaniappan [16] introduced modified invariant moments and proved that Hu [9] moments are not invariant for symmetric images. Further in his work it was proved that the new reference centre which was shifted to a distance from the image centroid remains invariant to the properties like scale, translation and rotation. It is also evident that the derived moment invariants show improvement to the symmetric images and are robust to noise variations. The characters in various font styles of English have several such characters, Hu's [9] moments are not suitable for this work; hence the new modified invariant moments suggested by Palaniappan have been used. Nagabhushan et. al [14] have used the modified invariant moments in recognition of script in pin code printed/handwritten numerals in the different Indian scripts.

In this paper we describe an approach based on scaling and translation invariant modified moments suggested in Palaniappan [16] for recognition of multi-font English characters. In section 2 proposed method is discussed. The section 3 deals with experiments and results and section 4 concludes the paper with conclusion.

2. Proposed Method

In the proposed method the isolated English character image is pre-processed for noise removal using median filtering followed with binarization using global threshold method. The pre-processed image is further inverted to get the background as black and objects as white. Region labeling is performed on the pre-processed image and a minimum rectangle bounding box is inserted over the character and the character is cropped. The cropped character is normalized to fit into a size of 33×33 pixel window without disturbing the aspect ratio of the character. Further on this character image thinning is performed using the thinning tool supported by Matlab 6.1. The end points of the character image are estimated. For end point determination zoning scheme is applied, the zone in which the end point lies is determined. The different zones used are depicted in fig. 2.

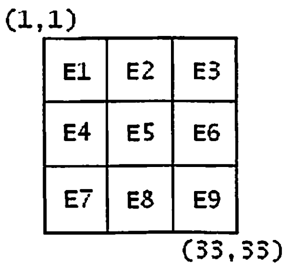

Fig. 2: Zones for end point classification

Further the aspect ratio and the Euler number of the isolated character image is estimated. Euler number of an image is defined as the number of objects minus number of holes. For example the characters 'A', 'D', 'O', 'R' etc. will have the Euler number 0 and the characters 'C', 'E', 'F', 'G' etc. will have Euler number 1. The characters are pre-classified initially in 4 groups according to their Euler number viz. 0, 1, -1 and 2. Further within these classes of the characters are divided according to their end points. Within these classes the characters the scale and translation invariant modified moments are estimated as shown in the section 2.1.

2.1. Modified Invariant Moments

The new modified invariant moments suggested by Palaniappan [16] are as follows. The new central moments are defined as

$$\lambda_{pq} = \sum_{x} \sum_{y} (x - \bar{x} + x_s)^p (y - \bar{y} + y_s)^q f(x, y)\, dx\, dy$$

Where p, q ε N_o as order indices, (x, y) are Cartesian co-ordinates, f is a non-negative continuous image function. $\bar{x}$ and $\bar{y}$ are means of image in x and y directions respectively and x_s and y_s are the modification factors. As the binary image functions are considered so f(x, y) can be only 0 or 1. The modified invariant moments are estimated by performing the calculations indicated as follows. The geometric moments are obtained using expressions (2.1)

$$\left.\begin{aligned} m_{00} &= \sum_{1}^{H}\sum_{1}^{W} f(x,y) \\ m_{01} &= \sum_{1}^{H}\sum_{1}^{W} y * f(x,y) \\ m_{10} &= \sum_{1}^{H}\sum_{1}^{W} x * f(x,y) \end{aligned}\right\} \quad (2.1)$$

where the terms H and W *are* the height and width of the image

The mean in x and y directions $\bar{x}$ and $\bar{y}$ are estimated using expression (2.2)

$$\bar{x} = \frac{m_{10}}{m_{00}}, \qquad \bar{y} = \frac{m_{01}}{m_{00}} \qquad (2.2)$$

The second order moments μ_{20} and μ_{02} are estimated using expressions (2.3).

$$\left.\begin{aligned} \mu_{20} &= \sum_{1}^{H}\sum_{1}^{W} (x - \bar{x})^2 f(x,y) \\ \mu_{02} &= \sum_{1}^{H}\sum_{1}^{W} (y - \bar{y})^2 f(x,y) \end{aligned}\right\} \quad (2.3)$$

These second order moments are used to find the modified invariant moments. To obtain the modified invariant moments, the modification factors x_s and y_s are estimated using the expression (2.4)

$$x_s = \sqrt{\frac{\mu_{20}}{m_{00}}}, \quad y_s = \sqrt{\frac{\mu_{02}}{m_{00}}} \qquad (2.4)$$

The normalized central moments are derived using the expression (2.5)

$$\phi_{pq} = \frac{\lambda_{pq}}{m_{00}^{(p+q+2)/2}} \qquad (2.5)$$

The set of 7 modified moment invariants are computed using following expressions

$$\varphi_1 = \phi_{20} + \phi_{02}$$

$$\varphi_2 = (\phi_{20} - \phi_{02})^2 + 4\phi_{11}^2$$

$$\varphi_3 = (\phi_{30} - 3\phi_{12})^2 + (3\phi_{21} - \phi_{03})^2$$

$$\varphi_4 = (\phi_{30} + \phi_{12})^2 + (\phi_{21} + \phi_{03})^2$$

$$\varphi_5 = (\phi_{30} - 3\phi_{12})(\phi_{30} + \phi_{12})\left[(\phi_{30} + \phi_{12})^2 - 3(\phi_{21} + \phi_{03})^2\right] + \ldots$$

$$\ldots(3\phi_{21} - \phi_{03})(\phi_{21} + \phi_{03})\left[3(\phi_{30} + \phi_{12})^2 - (\phi_{21} + \phi_{03})^2\right]$$

$$\varphi_6 = (\phi_{20} + \phi_{02})\left[(\phi_{30} + \phi_{12})^2 - (\phi_{21} + \phi_{03})^2\right] + \ldots$$

$$\ldots 4\phi_{11}(\phi_{30} + \phi_{12})(\phi_{21} + \phi_{03})$$

$$\varphi_7 = (3\phi_{21} - \phi_{03})(\phi_{30} + \phi_{12})\left[(\phi_{30} + \phi_{12})^2 - 3(\phi_{21} + \phi_{03})^2\right] + \ldots$$

$$\ldots(3\phi_{12} - \phi_{30})(\phi_{21} + \phi_{03})\left[3(\phi_{30} + \phi_{12})^2 - (\phi_{21} + \phi_{03})^2\right]$$

Thus the modified invariant moments estimated for each character class with different font-style and the mean set for each of the character class is determined and stored as the library. Initially the system was trained by taking one image per character per font style for seven fonts of the size 20. Thus the total training images were 364. The mean feature set for each character class is estimated and is stored in the feature set as library, thus treating this recognition problem as 52 class problem with one class per character. The character image under classification is pre-classified into one of the four classes based on the Euler number. Then within these classes it is assigned to one of the class with the aid of the end points. Further the modified invariant moments are estimated as shown in section 2.1. The character is assigned to a class with the minimum Euclidean distance mean vector. The nearest mean vector is determined using minimum distance nearest neighbor as follows.

Euclidean distance (ED) = $\sqrt{((x_1 - x_2)^2 + (y_1 - y_2)^2)}$ where (x_1, y_1) and the (x_2, y_2) are points in 2-D. The algorithm for the above procedure is presented in the section 2.2.

2.2. Algorithm: Recognition of Multi-font Characters

Input: Isolated Character image
Output: Recognition of the character
Method: Modified Invariant moments and minimum distance NN classifier.

Step1: Preprocess the image to eliminate noise using median filtering. Convert the pre-processed image in to binary form using global threshold method. Then invert the image so that the background is black and object is white.
Step2: Perform the region labeling for the pre-processed image, fit a minimum rectangle bounding box and crop the image.
Step3: Normalize the cropped image to a window of size 33×33 without disturbing the aspect ratio.
Step4: Perform the thinning operation on the normalized image using thinning tool of image processing tool box.
Step5: Compute the Euler number of the image and aspect ratio of the image.
Step6: Estimate the end points of the image character using zoning method
Step7: Compute the modified invariant moments of the image.
Step8: Pre-classify the image initially with the help of Euler number and end points.
Step9: Obtain the Euclidean distance between the invariant moments for test image and within class stored library by using the minimum distance nearest neighbor classifier.
Step10: Assign the image to a class with minimum distance.
Stop.

3. Experimental Results

For experimentation purpose 7 different font styles viz. Arial, Tahoma, Verdana, Lucida console, MS San Serif, Trebuchet and Century Gothic are considered. For training each character image for 7 different font styles and of size 20 are considered amounting to a total of 364 images. For implementation of the algorithm the character images with 20 different standard sizes of characters between 8 and 72 for the seven font styles were considered. Thus 20 images of each character class per font style constituting a total of 7,280 images were considered for testing. The method is implemented on Pentium IV system with 128 MB RAM machine with Matlab 6.1 software. The results of the classification for upper and lower case character images are presented in the Table 1 and Table 2 respectively.

Table 1: Results of upper case character recognition

Character	Tested Images	Correctly classified	% Accuracy
A	140	140	100
B	140	140	100
C	140	140	100
D	140	140	100
E	140	140	100
F	140	140	100
G	140	140	100
H	140	137	97.85
I	140	140	100
J	140	138	98.57
K	140	140	100
L	140	140	100
M	140	140	100
N	140	140	100
O	140	140	100
P	140	140	100
Q	140	140	100
R	140	140	100
S	140	140	100
T	140	140	100
U	140	140	100
V	140	140	100
W	140	140	100
X	140	136	97.14
Y	140	140	100
Z	140	140	100
Total	**3640**	**3631**	**99.75**

The Table 1 presents the classification of the upper case English characters for 7 different font styles. The recognition accuracy for the upper case character images is 100% except for four characters H, J, X and Z. The overall accuracy for upper case classification is 99.75% and is significant. The classification process does not distinguish the characters which are very similar in appearance in some font styles, eg. in Arial font style the lower case 'L' character and upper case 'I' character are similar in shape and size, even with human vision it is not possible to distinguish them. Hence it is treated as correct recognition.

The Table 2 presents the classification for the lowercase character images for 7 different font styles. The overall recognition accuracy for lower case character images is 99.56% and thus the overall accuracy is 99.65%.

Table 2: Results of lower-case character recognition

Character	Tested Images	Correctly classified	% Accuracy
a	140	140	100
b	140	140	100
c	140	140	100
d	140	140	100
e	140	138	98.57
f	140	140	100
g	140	138	98.57
h	140	140	100
i	140	140	100
j	140	140	100
k	140	140	100
l	140	140	100
m	140	140	100
n	140	140	100
o	140	140	100
p	140	140	100
q	140	138	98.57
r	140	140	100
s	140	140	100
t	140	138	98.57
u	140	140	100
v	140	140	100
w	140	138	98.57
x	140	136	97.14
y	140	140	100
z	140	138	98.57
Total	**3640**	**3624**	**99.56**

4. Conclusion

In this paper a method is proposed for Multi-font English character recognition which is based on modified invariant moments. It is also evident that the derived modified Invariant moment features are translation and size invariant. The novelty of this method is that it is simple, robust in the recognition of symmetrical images and showed encouraging results multi-font and multi-size English characters. The overall success rate of the algorithm is found to be 99.65%. The future work focuses on inclusion of as many number of font styles as possible.

5. References

[1] Abu Y. S. Mustafa and D. Psaltis, *"Recognition aspects of moment invariants"*, IEEE Trans. on PAMI, Vol. PAMI-6, pp. 698-706, Nov. 1984.

[2] Boyce J. F. and W. J. Hossach, *"Moment invariants for pattern recognition"*, Pattern Recognition Letters, Vol. 1 no. 5-6, pp. 451-456 1983.

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

[4] Cho-Huak Teh and Roland T. Chin, *"On Image Analysis by the methods of moments"*, IEEE Trans. on PAMI, Vol. 10 no. 4, pp. 496-512, 1988.

[5] Flusser J. and Tomas Suk, *"Affine moment Invariants: A new tool for character recognition"*, Pattern Recognition Letters, Vol. 15, pp. 16-22, 1994.

[6] Flusser J., *"Moment Invariants in Image Analysis"*, Trans. on Engineering, Computing and Technology, Vol. VII, pp. 196-201, Feb. 2006.

[7] Govindan V. K. and A.P. Shivaprasad, *"Character Recognition – a review"*, Pattern Recognition, Vol. 27(7), pp. 671-683, 1990.

[8] Hannan W. J., *"RCA multifont reading machine"*, Optical Character Recognition, eds. G. L. Ficher et al, McGregor and Wemer, pp 3-14, 1962.

[9] Hu M. K., *"Visual Pattern Recognition by moment invariants"*, IRE Trans. on Information theory, Vol. IT-8, pp. 179-187, 1962.

[10] Liao S.X. and Pawlak M., *"On Image Analysis by moments"*, IEEE Trans. on PAMI, Vol. 18, no. 3, pp. 254-266, 1996.

[11] Kahan S., T. Pavlidis and H. S. Baird, *"On the recognition of printed characters of any font & size"*, IEEE Trans. on PAMI, Vol. 9, no. 2, pp 274-287, 1987.

[12] Khotanzad A. and Y.H. Hong, *"invariant image recognition by Zernike moments"*, IEEE Trans. on PAMI, Vol. 12, no. 5, pp. 489-497, 1990.

[13] Mori S., C. Y. Suen and K. Yamamoto, *"Historical review of OCR research and Development"*, Proc. of IEEE-80, pp. 1029-1058, July 1992.

[14] Nagabhushan P., S. A. Angadi and B. S. Anami, *"An Intelligent Pin code script identification methodology based on texture analysis using modified invariant moments"*, Proc. of International Conference on Cognition &

Recognition (ICCR05), Mysore, India, Allied publishers, pp. 615-623, Dec. 2005.

[15] Nagy G., *"Twenty years of Document Image Analysis"*, IEEE Trans. on PAMI, Vol. 22, no. 1, pp. 38-62, 2000.

[16] Palaniappan R., P. Raveendran and Sigoru Omata, *"Improved moments invariant for invariant image representation"*, Invariants for Pattern recognition and Classification, World Scientific Publishing Co., pp. 167-185, 2000.

[17] Reddi S., *"Radial and angular moment invariants for Image identification"*, IEEE Trans. on PAMI-3, pp.240-242, 1986.

[18] Teague M. R., *"Image analysis via the general theory of moments"*, Jour. Optical Society America, Vol. 70, pp. 920-30, 1980.

[19] Trier O.D., A. K. Jain and T. Taxt, *"Feature extraction methods for character recognition- A survey"*, Pattern Recognition, Vol. 29, no. 4, pp. 641-662, 1996.

[20] Yang L. and F. Albregtsen, *"Fast computation of invariant moments: A new method giving correct results"*, Proc. of 12th International Conference on Pattern Recognition, Jerusalem, Israel, Vol. 1, pp. 201-204, Oct. 1994.