Multi-font/size Kannada Vowels and Numerals Recognition Based on Modified Invariant Moments

Mallikarjun Hangarge
Karnatak Arts, Science and Commerce Collage, Bidar, India

Shashikala Patil
Poojya Doddappa Appa college of Engineering, Gulbarga, India

B.V.Dhandra
Department of P.G. studies and Research in Computer Science, Gulbarga University, Gulbarga, India

ABSTRACT

In this paper, an attempt is made to develop an algorithm for recognition of machine printed isolated Kannada vowels and numerals of different font size and style using modified invariant moments and that are invariant with respect to rotation, scale and translation. A minimum distance nearest neighbor classifier is adopted for classification. The proposed algorithm is experimented on 1800 images of vowels and 1000 images of numerals. The experimental results confirm the recognition accuracy as of 97.7% for vowels and 98.92% for numerals. The algorithm is simple, robust and invariant with respect to rotation, scale and translation of an image.

General Terms:

Pattern Recognition, Document Image Analysis

Keywords: OCR, Modified Invariant Moments

1. INTRODUCTION

Optical character recognition (OCR) is one of the important active areas of research in pattern recognition. Presently, all communications, businesses and trades are performed through e-technology. Hence, the development of reliable OCR system is inevitable for different scripts and languages. Even though, reasonably good OCR systems are available in the market for local and universal languages and scripts, the rate of recognition drastically drops down when OCR reads a document containing different font style and font size of the characters. For example, newspaper containing (heading/subheadings), pamphlet with artistic characters and text books of kids etc. Therefore, bulk of research is going on for development of efficient and robust OCR system for different scripts and languages. In order to overcome such complexities, a number of methods have been proposed for the recognition of Indian scripts like Telugu, Tamil, Bangla, and Devanagari.

Ashwin *et al* [13] have formed the three basic Zones for the underlying character image. Each Zone is divided into a number of circular tracks and sectors. ON pixels in each angular region is used as a feature. Support vector machine was employed for the classification of characters and achieved an accuracy of 86.11%. A modified region decomposition method and optimal depth decision tree for the recognition of Kannada characters was used by Nagabhushan *et al* [15]. For the recognition of the characters of Indian scripts, entire image of the character was used as a feature vector as discussed in [16]. Sanjeev Kunte *et al* [17] have developed an OCR system for the recognition of basic characters of printed Kannada text, which works for different font size and font style. Each image was characterized by using Hu's invariant and Zernike moments. They have achieved the recognition accuracy as 96.8% with Neural Network classifier. Various types of moments have been used in the literature to recognize the image patterns in a number of applications. Moments considered include regular moments [1], Geometric [6, 5], Legendre moments [3], Zernike/pseudo-Zernike moments [7, 8, 10], radial and rotational moments [2, 12], and complex moments [4]. Recently, Palaniappan *et al* [9] introduced modified invariant moments for invariant image recognition and proved that Hu [11] moments are not robust to noise variation. 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. This survey of literature on Kannada OCR reveals that the research for Kannada character recognition is still in infant stage.

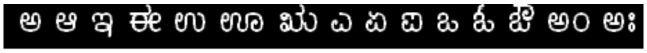
Therefore, in this paper, we have employed modified invariant moments as robust feature set for recognition of printed Kannada vowels and numerals of various font styles and sizes .This work is an extension of [20].

This paper is organized as follows. Section 2 of the paper contains the details of dataset used for experimentation and preprocessing methodology. The method of feature extraction is presented in Section 3 and the method of classification is given in Section 3.3. Experimental results and comparative analysis are the subject matter of Section 4. Section 5 contains conclusion and future plan of work.

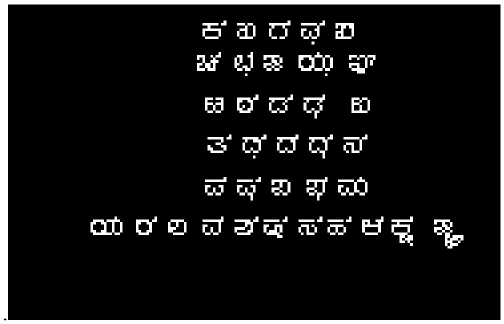
2. DATA SET AND PREPROCESSING

Kannada language is one of the fourth major south Indian languages spoken by about 50 million people. The Kannada script consists of 16 vowels, 36 consonants and 10 numerals. The sequence of Kannada characters is written horizontally from left to right to form a word and sentence. Kannada is a non-cursive script. The characters are isolated within a word. Most of the Kannada characters are circular in shape. A sample of vowels and numerals of different font styles and sizes are shown in Fig. 1 (a), (b), (c) and (d).

The standard database of Indian scripts is not available. Hence, we have created own database of Kannada characters using Nudi 4.0 and Baraha 8.0 software's. Different font styles like Nudi Akshara-01, Nudi Akshara-02 and Nudi Akshara-03 were used to create characters of size varying between 10 and 72 points. These are scanned through a flat bed HP scanner at 300 dpi. The noise is removed by using median filter. The gray scale images are then binarised using global threshold method [19] with object representing 1 and background of the object as 0. Then, the skew detection and correction of the scanned and binarized page is performed using algorithm [21]. Characters were segmented using horizontal and vertical projection profiles.

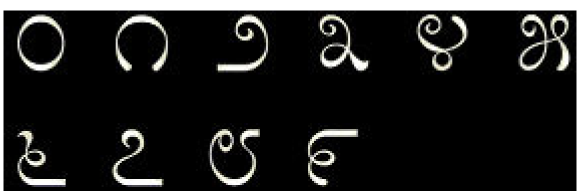


(a) Kannada Vowels of style Nudi Aksher 01 with regular font size 12



(b) Kannada Consonants of style Nudi Aksher 01 with regular font size 18

(c) Kannada vowels for different font style and size created using Nudi and Bhara software



(d) Kannada Numerals of style Nudi 03 k with size 30

Figure 1. Printed Kannada character set of different font style and size

3. FEATURE EXTRACTION

Region labeling is performed on the segmented characters and a minimum rectangle bounding box is inserted over the character. Scale, rotation and translation invariant modified moments suggested by Palaniappan [9] were computed and stored as the potential features vector. Brief descriptions about the modified invariant moments are given in the Section 3.1 that follows.

3.1 Modified Invariant Moments

The new modified invariant moments suggested by Palaniappan [9] are as follows. The new central moments with modification factors x_s and y_s are defined as

$$\lambda_{pq} = \sum_x \sum_y (x - \bar{x} + x_s)^p (y - \bar{y} + y_s)^q f(x,y)dxdy$$

where p and q as order indices, (x, y) is Cartesian coordinates, 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 , y_s are the modification factors. Since the image functions f(x, y) can be only 0 or 1. The modified invariant moments are computed by using the following formulae.

$$m_{00} = \sum_1^H \sum_1^W f(x,y)$$

$$m01 = \sum_1^H \sum_1^W y * f(x,y) \qquad (2.1)$$

$$m_{10} = \sum_1^H \sum_1^W x * f(x,y)$$

where H is height of the image and w is width of the image the mean $\bar{x}$ and $\bar{y}$ are obtained by using expression.

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

The second order moments m_{20} and m_{02} are estimated using the expressions

$$m_{20} = \sum_1^H \sum_1^W \left(x - \bar{x}\right)^2 f(x,y)$$

$$m_{02} = \sum_1^H \sum_1^W \left(y - \bar{y}\right)^2 f(x,y) \qquad (2.3)$$

where H and W are as defined in (2.1) .The modification factors x_s and y_s present in modified invariants are estimated using the following expression:

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

The normalized central moments are derived using the expression.

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

The set of 7 modified invariant moments are computed by using following expressions.

$$\mu_1 = \phi_{20} + \phi_{02}$$

$$\mu_2 = (\phi_{20} - \phi_{02})^2 + 4\phi_{11}^2$$
$$\mu_3 = (\phi_{30} - 3\phi_{12})^2 + (3\phi_{21} - \phi_{03})^2$$
$$\mu_4 = (\phi_{30} + \phi_{12})^2 + (\phi_{21} + \phi_{03})^2$$
$$\mu_5 = (\phi_{30} - 3\phi_{12})(\phi_{30} + \phi_{12})\left[(\phi_{30} + \phi_{12})^2 - 3(\phi_{21} + \phi_{03})^2\right] +$$
$$..(3\phi_{21} - \phi_{03})(\phi_{21} + \phi_{03})\left[3(\phi_{30} + \phi_{12})^2 \_ (\phi_{21} + \phi_{03})^2\right]$$
$$\mu_6 = (\phi_{20} + \phi_{02})\left[(\phi_{30} + \phi_{12})^2 - (\phi_{21} + \phi_{03})^2\right] +$$
$$4\phi_{11}(\phi_{30} + \phi_{12})(\phi_{21} + \phi_{03})$$
$$\mu_7 = (3\phi_{21} - \phi_{03})(\phi_{30} + \phi_{12})\left[(\phi_{30} + \phi_{12})^2 - 3(\phi_{21} + \phi_{03})^2\right]..+$$
$$+(3\phi_{12} - \phi_{30})(\phi_{21} + \phi_{03})\left[3(\phi_{30} + \phi_{12})^2 - (\phi_{21} + \phi_{03})^2\right]$$

The modified invariant moments are computed for each character of different font-style and font size and the mean of moments of all character is determined and stored as feature vector.

3.2. Algorithm: Feature Extraction

Input: Isolated Kannada vowels and numerals image.

Output: Feature library.

Begin

1. Convert the image into binary form using global threshold method. Then invert the image such that the background is black and object is white
2. Perform the region labeling for the preprocessed image f, fit a minimum rectangle-bounding box on it.
3. Compute the modified invariant moments on f.
4. Repeat the step 1 to step 3 for all sample images representing vowels and numerals considered for training
5. Compute mean and store the computed invariant moments as the feature vector.
6. Repeat above steps for all vowels and numerals

End

3.3 Classification

Euclidean distances between the invariant moments of training images and test images were obtained and the nearest-neighbor classifier is used for recognition of test character images. The nearest-neighbor classifier relies on distance function between patterns. The Euclidean distances between the invariant moments of the trained sample with that of test sample is obtained by using the following formula:

$$\text{d (a,b)} = \sqrt{\sum_{i=1}^{7} (a-b)}^{\,2}$$

where ***a*** is the mean feature vector of the trained image and ***b*** is the feature vector of the test character image. The test image is preprocessed and modified invariant moments were computed as described in Section 3.1. The Euclidean distances between the training feature vector and testing feature vector is calculated and stored in the library. The minimum distance is determined and test image is declared to belong to the class, which is presented in the Section 4.1.

3.4 Algorithm: Character Recognition

Input: Isolated testing vowel and numeral images.

Output: Recognition of printed Kannada vowels and numerals

Begin

1. Call feature extraction algorithm presented in 3.2
2. Compute the Euclidean distance between the moments on testing images with that of training feature vectors of each class stored in the library.
3. Assign the image to a class having minimum distance.

End.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

For verification and validation of the algorithm1800 vowels and 1000 numerals of different font style and size of Kannada characters were used. Among 2800 images, we have randomly divided into two sets. One set is for training and another for testing. The classifications results obtained after conducting the experimentation on Kannada vowels and numerals are presented in Table 1 and 2. These results reveal that the modified invariant moments with basic NN classifier is a robust procedure with respect to multi font styles and sizes. Hoverer, with other classifiers like KNN and SVM, this algorithm may give possibly the best results. The average percentage of recognition accuracy of the proposed method is 97.7% and 98.92%. A comparative study of the proposed method with existing methods proposed by Ashwin and Sastry [13] and Sanjeev Kunte[17] is presented in Table 3. However, based on this comparison, the robustness of the proposed method cannot be justified, because of the dataset, experimental conditions, number of features and classifiers used by the other methods were different. The method is implemented on Pentium IV System with 512 MB RAM with Matlab 7.0

Table 1. Kannada Vowels Recognition Accuracy with NN

Vowels	Training Images	Testing Images	Accuracy in %
ಅ	150	150	100
ಆ	150	142	94.66

ಇ	150	150	100
ಈ	150	150	100
ಉ	150	145	96.66
ಊ	150	146	97.33
ಎ	150	150	100
ಏ	150	144	96
ಐ	150	146	97.33
ಒ	150	150	100
ಓ	150	142	100
ಔ	150	144	96
Total	1800	1759	97.7

Table 2. Kannada Numerals Recognition Accuracy with NN

Numerals	Training Images	Testing Images	Accuracy in %
೦	50	50	100
೧	50	50	99.81
೨	50	50	98.55
೩	50	50	98.19
೪	50	50	99.17
೫	50	50	100
೬	50	50	100
೭	50	50	97.44
೮	50	50	99.41
೯	50	50	96.94
Total	500	500	98.92%

5. CONCLUSION.

In this paper, simple and robust method for multi-font/size printed Kannada vowels and numerals recognition is proposed based on modified invariant moments. The features of derived and modified moments are invariant with respect to translation, rotation and size. The experimental results presented above exhibits that the efficiency of the modified invariant moments is higher than other methods stated. The novelty of this method is its simplicity. It is free from thinning and normalization of the characters. In future, effort will be made to improve the algorithm for complete Kannada OCR system in such a way that it should be efficient to read poor documents (printing and paper quality), degraded and artistic documents etc. Further, we are intended to extend it for designing south Indian multi-script character recognition system.

Table 3 Comparison Analysis

Name	Features Employed	Classifier used	Character Recognition Rate
Ashwin and Sastry[13]	Zoning features.	Support Vector machine.	86.11%
Sanjeev Kunte[17]	Zernike features	RBF neural classifier	96.80%
Proposed method	Modified invariant moments	Nearest neighbor classifier.	97.34%

6. REFERENCE.

[1] Liao S.X. and Pawlak M., "On image analysis by moments invariants", IEEE Trans. On PAMI, vol.18, no. 3, pp. 254-266, 1996.

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

[3] Cho-Huak the and Roland T. Chin, "On image Analysis by the methods of moments", IEEE Trans on PAMI, VOl.10 no. pp.496-512, 1988.

[4] Abu Y. S. Mustafa and D. Psaltis, "Recognition aspects of moment invariants", IEEE Trans. On PAMI, Vol. PAMI-6, pp.698-706, NOV.1984.

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

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

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

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

[9] 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.

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

[11] Hu M. K., "visual pattern Recognition by moment invariants", IRE Trans. On Information theory, Vol.IT-8, pp. 179-187, 1962.

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

[13] Ashwin T v, Sastry P S 2002 A fonts and size-independent OCR system for printed Kannada documents using support vector machines. Sadhana 27:35-58

[14] Negi Atual, Chakravarthy Bhagavathi, Krishna B 2001 An OCR system for Telugu. Proc. Sixth Inter. Conference. On Document Anal. And rec. 1110-1114

[15] Nagbhushan P, Pai Radhika M 1999 Modified region decomposition method and optimal depth decomposition tree in the recognition of non –uniform sized characters – An experimentation with Kannada characters. Pattern Recognition .Letter. 20:1467-1475

[16] Jawahar C V, Pavan Kumar , Ravi Kiran S S 2003 Bilingual OCR for Hindi & Telugu documents & its Applications. Proc. Seventh Int.Confer. On Documents Anal ana Rec. 408-412.

[17] Kunte Sanjeev R, Sudhaker Samuel (2006), Hu's invariant moments & Zernike moments approach for the recognition of basic symbols in printed Kannada text. Sadhana vol .32, part 5, October 2007, pp. 521-533.

[18] Chaudhuri B.B. and U.Pal, An OCR system to read two Indian language scripts: Bangle and Devanagari, Proceedings of ICDAR, 1997, 1011-1015.

[19] R.C.Gonzal, R.E. Woods, "Digital Image Processing", Pearson Education, 2002.

[20] B.V.Dhandra, Mallikarjun Hangarge and shashikala Patil 2010, Multi-Font Kannada Vowels Recognition Based on Modified Invariant Moments. Proceedings of RTIPPR-2010, 119-122.

[21] Skew detection in Binary image documents based on Image Dilation and Region labeling Approach, Proceedings of 18th International Conference on Pattern Recognition - ICPR 2006, Hong Kong 2006, V. No. II-3, pp. 954-957