ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 481– 485 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

Feature Extraction Techniques for Recognition of Malayalam Handwritten Characters: Review

Ashlin Deepa R.N[1], R.Rajeswara Rao[2]
1 Assistant Professor, Gokaraju Rangaraju Institute of Engineering and Technology,Bachupally, Kukatpally,
Hyderabad,
Andhra Pradesh,India.- 500090
deepa.ashlin@gmail.com
2 Associate Professor of Computer Science & Engineering,
JNTUK University College of Engineering,
Vizianagaram,
JNTU Kakinada, Andhra Pradesh, India.
raob4u@yahoo.com

Abstract: The Character recognition is one of the most important areas in the field of pattern recognition. Recently Indian Handwritten character recognition is getting much more attention and researchers are contributing a lot in this field. But Malayalam, a South Indian language has very less works in this area and needs further attention. Malayalam OCR is a complex task owing to the various character scripts available and more importantly the difference in ways in which the characters are written. The dimensions are never the same and may be never mapped on to a square grid unlike English characters. Selection of a feature extraction method is the most important factor in achieving high recognition performance in character recognition systems. Different feature extraction methods are designed for different representation of characters. As an important component of pattern recognition, feature extraction has been paid close attention by many scholars, and currently has become one of the research hot spots in the field of pattern recognition. This article gives a general discussion of feature extraction techniques used in handwritten character recognition of other Indian languages and some of them are implemented for Malayalam handwritten characters.

Keywords: ***character recognition, pattern recognition, feature extraction.***

INTRODUCTION

Character recognition is one of the most interesting and challenging research areas in the field of Image processing. Nowadays different methodologies are in widespread use for character recognition. Various approaches of hand written character recognition are discussed here along with their performance. It can contribute immensely to the advancement of an automation process and can improve the interface between man and machine. It is the mechanism to convert machine printed, hand printed or hand written document file into editable text format. Major Steps of an OCR System are described in fig.1.

FEATURE EXTRACTION

Feature extraction is finding the set of parameters that define the shape of a character precisely and uniquely. Selection of feature extraction method is probably one of the most important characters for achieving high performance.

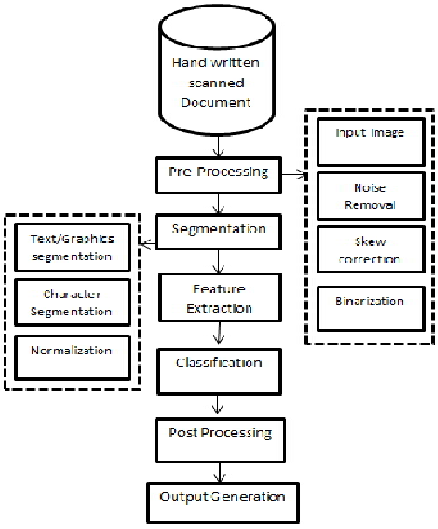


Fig.1 OCR System

Feature extraction [2] methods are classified into three major groups as:

Statistical features.
Global transformation and series expansion
Geometric and topological features

Statistical features

Statistical features represent the image as statistical distribution of points. Statistical features are derived from the statistical distribution of points. They provide high speed and low complexity and take care of style variations to some extent. They may also be used for reducing the dimension of the feature set. Various methods which use statistical features are Zoning, Crossings and Distances, Projections etc.

1) Zoning

The statistical feature adopted in this research is 'Zoning'. Zone-based feature extraction method provides good result even when certain pre processing steps like filtering,

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 481– 485 (2014)

Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

smoothing and slant removing are not considered [4]. The commercial OCR system by Calera described in Bokser [9] uses zoning on solid binary characters. A straightforward generalization of this method to gray level character images is given here.

The same method is applied on Malayalam character (DA). An nm grid is superimposed on the character image (DA) (Fig. 2(a)), and for each of the nm zones, the average gray level is computed (Fig. 2(b)), giving a feature vector of length n x m. However, these features are not illumination invariant.

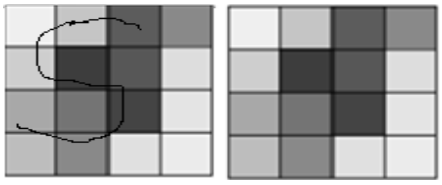


Fig.2 (a) The zoning of grey level character image (b) Average grey levels in each zone

The frame containing the character is divided into several overlapping or non- overlapping zones and the densities of object pixels in each zone are calculated. Density is calculated by finding the number of object pixels in each zone and dividing it by total number of pixels [3]. Image Centroid and zone-based (ICZ) distance metric feature extraction and Zone Centroid and zone-based (ZCZ) distance metric feature extraction algorithms were proposed by Vanajah and Rajashekararadhya in 2008 for the recognition of four popular Indian scripts (Kannada, Telugu, Tamil and Malaya- lam) numerals. In this research, hybrid of modified Image Centroid and zone-based (ICZ) distance metric feature extraction and modified Zone Centroid and zone-based (ZCZ) distance metric feature extraction methods was used. Modifications of the two algorithms are in terms of:

(i) Number of zones being used

(ii) Measurement of the distances from both the Image Centroid and Zone Centroid

(iii) The area of application.

The image (Malayalam character) is further divided in to 'n' equal parts (Twenty five in this case) as shown in Figure 2. The character centroid (i.e. centre of gravity of the character) is computed and the average distance from the character centroid to each pixel present in the zone is computed. Similarly zone centroid is computed and aver- age distance from the zone centroid to each pixel present in the zone is to be computed. This procedure will be repeated for all the zones/grids/boxes present in the character image. There could be some zones that are empty, and then the value of that particular zone image value in the feature vector is zero. Finally, 2 x 25 (i.e. Fifty in this case) such features were used to represent the character image feature. Fig.3 and Fig 4 present the method applied for Malayalam character image.

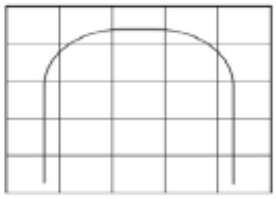


Fig.3 Malayalam Character 'RA' in 5 y 5(25 equal zones)

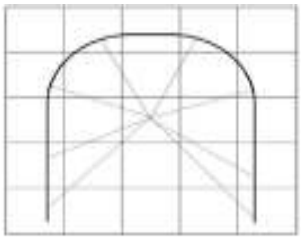
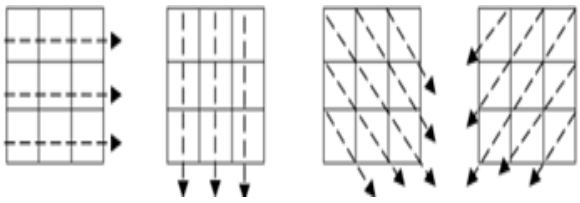


Fig.4 Image centroid on 'RA' in zoning

2) Projection Histogram Features

Projection histograms count the number of pixels in specified direction. In paper [4], this approach is applied and three directions of horizontal, vertical and diagonal traversing is done. Three types of projection histograms horizontal, vertical, diagonal-1 (left diagonal) and diagonal-2 (right diagonal) were created. These projection histograms for a 3*3 pattern are depicted in Fig. 5. In this approach projection histograms are computed by counting the number of foreground pixels. In horizontal histogram these pixels are counted by row wise i.e. for each row. In vertical histogram the pixels are counted by column wise. In diagonal-1 histogram the pixels are counted by left diagonal wise. In diagonal-2 histogram the pixels are counted by right diagonal wise. The lengths of these features are 32, 32, 63 and 63 respectively according to lines of traversing.



(a) (b) (c) (d)

(a) Horizontal Histogram (b) Vertical Histogram (c) Diagonal-1 Histogram (d) Diagonal-2 Histogram

Fig.5 Evaluation of 4 types of Projection Histograms on 3*3 patterns

Glauberman [6] used projection histograms first time in a hardware based OCR. This technique is generally used to detect orientation in a document page or segmenting a page into lines, words and characters. In order to find the projection histograms, an image is tracked along a path from a side and the number of black pixels in that path is counted. A histogram gives the width of character strokes along a particular path (either row or column). The vertical histograms are slant invariant whereas horizontal histograms are not [2]. The histograms only give the stroke information along the given path and do not cover any other

ISSN 2278-3091
International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 481– 485 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

properties such as number of strokes along a path, width of each stroke, location of each stroke, etc

3) Distance Profile Features

Satish kumar [5] used Profile which counts the number of pixels (distance) from bounding box of character image to outer edge of character. This traced distance can be horizontal, vertical or radial. The approach used in this paper is profiles of four sides left, right, top and bottom. Left and right profiles are traced by horizontal traversing of distance from left bounding box in forward direction and from right bounding box in backward direction respectively to outer edges of character. Similarly, top and bottom profiles are traced by vertical traversing of distance from top bounding box in downward direction and from bottom bounding box in upward direction respectively to outer edges of character. The size of each profile in this approach is 32.

4) Crossings

Crossings based methods have been used in [7], [8] for hand-printed character recognition and are generally used to detect the number of strokes presented in a character along a particular path. If this path is along the rows, then it is called horizontal crossings and if this path is along the columns, then it is called vertical crossings. The crossings can be considered as the number of transitions either from black to white or from white to black pixels along a particular path. Crossings can be extracted from an original character as well as from its skeleton. Kim et al [7] used crossings in raw form but Arica et al [8] used median of the black pixel runs in each scan line. The region in which the median of a black pixel run falls is assign its code value otherwise the code value for the region is 0. The sum of codes due to all regions in a scan line is taken as feature.

Global transformation and series expansion

In global transformation and series expansion various techniques are :

1) Fourier transform
2) Gabor transform
3) Fourier Descriptor
4) Wavelets
5) Moments
6) Karhunen-Loeve expansion etc.

1) Fourier Descriptors

Shape feature vector consists of the Fourier descriptors. After the boundary pixel set of an object was computed. The method uses centroid distance function to compute shape signature from boundary pixels of a shape in a local space. Here this centroid distance function is the periodic function we consider and decompose into fourier series. A pair of shape signature and boundary pixel gray was used as a point in a feature space. Fourier transform is used for shape signature to compute Fourier coefficients, and standardized pixel brightness is introduced into computational process of the Fourier coefficients so that shape features can be computed. The Fourier coefficients which are invariant to translation, scaling, rotation and change of start point are used as Fourier descriptors [10].

Note: Fourier series means decomposing a periodic function into sum of set of sine and cosine functions. The coefficients corresponding to are the fourier coefficients.

- If we want invariance to translation, do not use the DC-term, that is the first element in your resulting array of fourier coefficients f[0].
- If we want invariance to scaling, make the comparison ratio-like, for example by dividing every Fourier coefficient by the DC-coefficient.

f*[1]= f[1]/f[0] and so on.

- If we want invariance to the start point of your contour, only use absolute values of the resulting fourier coefficients.

From these shape features are extracted. The method may also use complex coordinates and curvature function as shape signature.

2) Wavelets

Wavelets Transform represents a mathematical way used to study non-stationary signals. Therefore, its usefulness has been increasingly adapted over the last 10 years. It was employed in different fields such as communication technology, geophysics and image processing. The wavelet transform provides an appropriate basis for image handling because of its beneficial features. The assets of the wavelet transform are: The ability to compact most of the signal's energy into a few transformation coefficients, which is called "energy compaction" and the ability to capture and represent effectively low frequency components (such as image backgrounds) as well as high frequency transients (such as image edges). Wavelet transform coefficients are energy, variance and waveform length. The features are extracted from these coefficients. The use of discrete wavelet transform (DWT) both for signal preprocessing and signal segments feature extraction as an alternative to the commonly used discrete Fourier transform (DFT). Feature vectors belonging to separate signal segments are then classified by a competitive neural network as one of the methods of cluster analysis and processing.

By means of wavelet analysis, a matrix of data is obtained, where time and frequency domain information is present. Another waveform is ''compressed'' or ''stretched'' to obtain wavelets of different scales that are used along time comparing them with the original signal [11].

3) Moments

Moment based features are very effective in describing shape of characters. It is observed that moment based features can become very effective if certain operations such as normalization of character size and geometric operations are performed correctly using floating point arithmetic. we use the features drawn by invariants moment technique which is

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 481– 485 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

used to evaluate seven distributed parameter of a character image. The moment invariants (IMs) are well known to be invariant under translation, scaling, rotation and reflection. They are measures of the pixel distribution around the centre of gravity of the character and allow to capture the global character shape information. In the paper[12], the moment invariants are evaluated using central moments of the image function f(x,y)) up to third order. Moments constitute an important feature extraction method (FEM) which generates high discriminative features, able to capture the particular characteristics of the described pattern. Among the several moment families introduced in the past, the orthogonal moments are the most popular moments widely used in many applications, owing to their orthogonality property [13].

Geometric and topological features

The paper [14] describes a geometry based technique for feature extraction applicable to segmentation-based character recognition systems. This method extracts the geo-metric features of the character contour. These features are based on the basic line types that form the character skeleton. The system gives a feature vector as its out- put. The method is implemented for Malayalam characters and the various steps involved in geometric method are:

(i)Initially preprocessing (binarization, skeletonization) is done on the input image.

(ii)Universe of discourse is defined as the shortest matrix that fits the entire character skeleton. The Universe of discourse is selected because the features extracted from the character image include the positions of different line segments in the character image.

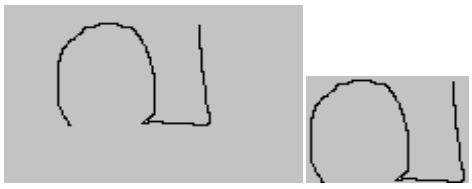


Fig. 6.(a) Original Image (b) Universe of Discourse

Malayalam character image 'VA' is given in Fig. 6(a) and Fig.6(b) gives its universe of discourse

iii) After the universe of discourse is selected, the image is divided into windows of equal size, and the feature is done on individual windows.

(iv)To extract different line segments in a particular zone, the entire skeleton in that zone should be traversed. For this purpose, certain pixels in the character skeleton were defined as starters, intersections and minor starters.

(v) After the line type of each segment is determined, feature vector is formed based on this information. Every zone has a feature vector corresponding to it. Under the algorithm proposed, every zone has a feature vector with a length of 8. The con- tents of each zone feature vector are

1) Number of horizontal lines
2) Number of vertical lines
3) Number of Right diagonal lines
4) Number of Left diagonal lines
5) Normalized Length of all horizontal lines.
6) Normalized Length of all vertical lines.
7) Normalized Length of all right diagonal lines
8) Normalized Length of all left diagonal lines
9) Normalized Area of the Skeleton

The number of any particular line type is normalized using the following method, value = 1 - ((number of lines/10) x 2)

Normalized length of any particular line type is found using the following method,

length = (Total Pixels in that line type)/ (Total zone pixels)

The feature vector explained here is extracted individually for each zone. So if there are N zones, there will be 9N elements in feature vector for each zone.

Table 1. Merits and demerits of three categories of feature extraction techniques

Feature Extraction Method	Merits	Demerits
Statistical Feature Extraction	Pattern from different classes are well separated and produces compact pattern set.	If we have a restricted amount of information it is sufficient for a direct solution but is insufficient for solving a more general
Global transformation and series expansion	Features give well representation and well normalization of shapes.	For more complex character shapes high frequency information is needed.
Geometric and topological Features	Relationship between the components in the shape is highly expressed.	Non-planar feature geometry continues to pose challenges for recognition techniques. Feature mapping from one application domain to another is another challenging problem. problem.

CONCLUSION

In this paper the feature extraction methods which are used for other Indian languages and can be used in Malayalam handwritten characters are discussed. Some of the techniques are applied on Malayalam dataset and the features are extracted. One of the major difficulties in this field is the lack of bench mark database for hand written characters for most of the languages for testing of research results.

REFERENCES

[1] Nisha Sharma, Tushar Patnaik, Bhupendra Kumar, " *Recognition for Handwritten English Letters: A Review*", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013.

[2] Ivind Due Trier, Anil K. Jain, and Torfinn Taxt, "*Feature Extraction methods for character recognition: A survey*", Department of

Computer Science, Michigan State University, A714 Wells Hall, East Lansing, MI 48824{1027, USA Revised July 19}, 1995.

[3] Pritpal Singh, Sumit Budhiraja, *" Feature Extraction and Classification Techniques in O.C.R.Systems for Handwritten Gurumukhi Script – A Survey"*, International Journal of Engineering Research and Applications, (IJERA) ISSN: 2248-9622, Available: www.ijera.com Vol. 1, Issue 4, pp. 1736-1739

[4] Kartar Singh Siddharth , Mahesh Jangid, Renu Dhir, Rajneesh Rani, "*Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features"*, Innovative Systems Design and Engineering, ISSN 2222-2871 (Online) Vol 3, No 3, 2012, Available: www.iiste.org

[5] Satish Kumar, "*Neighborhood Pixels Weights-A New Feature Extractor"*, International Journal of Computer Theory and Engineering, Vol. 2, No. 1 February, 2010 1793-8201

[6] M. H. Glauberman, "*Character Recognition for Business Machines*", Electronics 29, 1996, pp.132-136

[7] K. M. Kim, J.J. Park, Y.G. Song, I. C. Kim and C. Y. Suen, "*Recognition of Handwritten Numerals Using a Combined Classifier with Hybrid Features*", SSPR & SPR, LNCS 3138, 2004, pp. 992-1000.

[8] N. Arica and F. T. Yarman-Vural, "*Optical Character Recognition for Cursive Handwrit ing*", IEEE Transactions on Pattern Analysis and Machine Intelligence,2002, vol. 24, no. 6.

[9] M. Bokser, \Omni document technologies," Proceedings of the IEEE, vol. 80, pp. 1066-1078, July 1992

[10] Gang Zhang ; Coll. of Inf. Sci. & Eng., Northeastern Univ., Shenyang ; Ma, Z.M. ; Qiang Tong ; Ying, " *The shape feature extraction using fourier desciptors with Brightness*", presented at the, IIHMSP '08 International Conference on Intelligent Information Hiding and Multimedia Signal Processing,2008, Page(s):71 - 74 Print ISBN:978-0-7695-3278-3

[11] Kheder G., Kachouri A., Taleb R., Ben Messaoud M. and Samet M., "*Feature extraction by wavelet transforms to analyze the heart rate variability during two meditation technique"*, presented at the 6th WSEAS International Conference on Circuits, Systems, Electronics, Control & Signal processing, Cairo, Egypt, Dec 29-31, 2007.

[12] R. J. Ramteke, "*Invariant Moments Based Feature Extraction for Handwritten Devanagari Vowels Recognition*", 2010 International Journal of Computer Applications (0975 - 8887) Vo- lume 1 – No. 18.

[13] G.A. Papakostas, D.E. Koulouriotis and V.D. "*Tourassis Feature Extraction Based on Wavelet Moments and Moment Invariants in Machine Vision* " , Available: www.intechopen.com.

[14] Dinesh Dileep,*" A Feature Extraction Technique Based on Character Geometry for Character Recog*nition", Arxiv, 2012.