Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Study of Different Feature Extraction Techniques for Devnagari Handwritten Word Recognition

Saniya Ansari*
Research Scholar, Dept. of ECE,
Karpagam University, Coimbatore,
Tamil Nadu, India

Dr. Udaysingh Sutar
Professor, Dept. of ECE,
Karpagam University, Coimbatore,
Tamil Nadu, India

Abstract— In this paper Feature extraction is a very crucial step as the success rate of a recognition system is often attributed to a good feature extraction method. The feature extractor determines which properties of the pre-processed data are most meaningful and should be used in further stages. For online recognition Vertical position of a point, curvature, Pen Up/Pen Down, writing direction, aspect, and slope are amongst the various features extracted. The most important aspect of handwriting recognition scheme is the selection of good feature set & maximum feature vector size. The feature set should be reasonably invariant with respect to shape variations caused by various writing styles. In the proposed work we have used four types of feature extractor namely Regional features, gradient features, Structural features & distant transform features. The size of Feature vector is selected as 91 to achieve maximum accuracy.

Keywords—Gradient Feature Extractor, Distance Transform, Zoning, Regional features, Geometric Features, Euler Number

I. INTRODUCTION

Online handwritten character recognition is having wide areas of application in real life environment. Therefore the accuracy of such systems should be more, efficient and faster to process applications. A lots of research work is still going on over handwritten character recognition based on different languages and scripts. For any handwritten character recognition, there are three main tasks such as Image segmentation, Feature Extraction and Classification. Feature extraction is a very essential step for online handwritten character recognition. As the success rate of a recognition system is often depends on a good feature extraction method.

Feature extraction is also called as data extraction & gives data from perspective areas. Features are a set of numbers that capture the salient characteristics of the segmented image. The feature extraction methods for handwritten character recognition are based on two types of features: statistical and structural. The statistical features are derived from the statistical distributions of pixels, such as zoning, moments, projection histograms or direction histograms. Structural features are based on the topological and geometrical properties of the character, like strokes and their directions, end-points or intersection of segments and loops.

The widely used feature extraction methods are Template matching, Deformable templates, Unitary Image transforms, Graph description, Projection Histograms, Contour profiles, Zoning, Geometric moment invariants, Zernike Moments, Spline curve approximation, Fourier descriptors, Gabor feature. Due to the nature of handwriting with its high degree of inconsistency and ambiguity extracting these features, is a difficult task.

II. LITERATURE SURVEY

In [1], Satish Kumar observed that Kirsch directional edges are least performing and gradient is good performing with SVM classifiers. With MLP, the performance of gradient and directional distance distribution is almost same. The chain code based feature is better as compared to Kirsch directional edges and distance transform. In overall, the gradient based feature is better than others on Devanagari.

In [2], J. Pradeep, presented a handwritten character recognition system using multilayer Feed forward neural network. Three different orientations, namely, horizontal, vertical and diagonal directions are used for extracting 54 features from each character. The diagonal orientation for feature extraction is identified to be the most suitable method as it yields higher recognition accuracy. From the test results the diagonal method of feature extraction yields the highest recognition accuracy of 98% for 54 features and 99% for 69 features.

In [3], Brijmohan Singh, used two different methods for extracting features from handwritten Devnagari characters, the Curvelet Transform and the Character Geometry, and compare their recognition performances using two different classifiers, viz., the Support Vector Machine (SVM) with Radial Basis Function (RBF), and the k-Nearest Neighbor (k-NN) classifier. Results obtained show that Curvelet features with k-NN classifier performs the best, yielding accuracy as high as 93.8%.

© 2015, IJARCSSE All Rights Reserved

In [4], Vandita Singh presents a survey of techniques for recognition of handwritten and hand printed documents in off-line mode. Statistical techniques such as multi zoning and edge maps yielded very good efficiencies for hand written Roman alphabet recognition. For classification, the best efficiencies have been obtained by using SVM as the classifier.

In [5], Gita Sinha, presented an overview of Feature Extraction techniques for off-line recognition of isolated Devanagari numeral recognition. Zone based approach presents the combination of image centroid zone and zone centroid zone of numeral/character image. The 99.11% recognition accuracy is obtained by SVM.

In [6], Pratibha Singh, Features are calculated on three different zoning methods. Directional feature is considered which is obtained using chain code and gradient direction quantization of the orientations. For classification 1-nearest neighbor based classifier, quadratic bayes classifier and linear bayes classifier are chosen as base classifier. The base classifiers are combined using four decision combination rules namely maximum, Median, Average and Majority Voting. The framework is used to test the reliability of recognition system against ambiguity.

In [7], Munish Kumar presented a novel feature extraction technique for an offline handwritten Gurmukhi character recognition system. They have extracted various topological features, namely, peak extent features, shadow features and centroid features. A new feature set is also proposed by using horizontal peak extent features and the vertical peak extent features. For classification, they used k-NN and Linear-SVM classifiers. Proposed system achieves a maximum recognition accuracy of 95.62% using SVM with linear kernel classifier. By using k-NN and MLPs, a maximum recognition accuracy of 95.48% and 94.74%, respectively.

III. PROPOSED METHOD

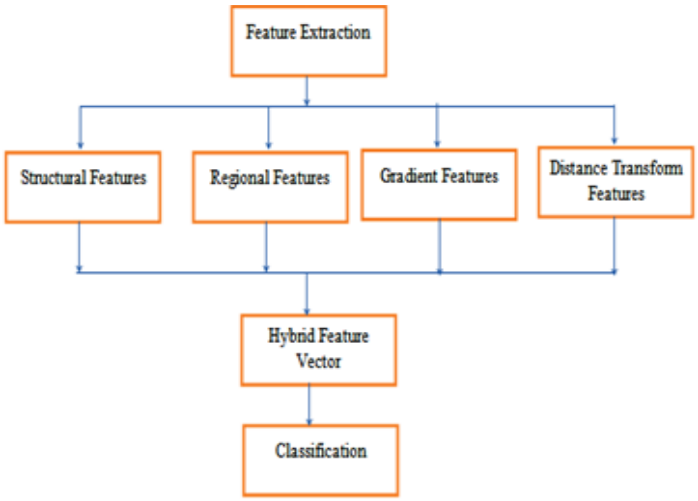

Fig. 3.1. Proposed Feature Extraction Method

As existing feature extraction methods are having limitations related to accuracy and time, in this paper we are presenting faster, efficient and optimized feature vector for handwritten character recognition. Below figure 3.1 is showing the block diagram of proposed feature vector methods used in handwritten character recognition.

First step of handwritten recognition is pre-processing and segmentation. Pre-processing is done for noise removal, resizing and conversion into image which is suitable for segmentation. After that segmentation is perform to extract the real information data from input image in order to do further processing. The second step of handwritten recognition is feature extraction. The selection of feature extraction technique is main factor for delivering highest handwritten recognition accuracy. There are many methods for feature extraction presented in literature. The most commonly used feature extraction methods are Gradient features, structural features, regional features, projection histograms, Zernike moments, zoning etc. Many of methods may require more time for feature extraction, however delivering better accuracy. But in handwritten character recognition, we require both high accuracy and less time.

In this work we are presenting new hybrid method of feature extraction in which total 91 features are used for recognition. These features are combination of structural or geometric features, regional features; gradient features and distance transform features. In existing case, combination of structural features, regional features and moment features were used, however there is more time required for structural features extraction. In this paper we optimized structural features by using Universe of discourse over input segmented image, which can speed up the tasks of 81 feature extraction. In addition to this, proposed feature vector delivers better accuracy as compared to existing method.

The feature extractor determines which properties of the preprocessed data are most significant and should be used in further stages. In this paper we are discussing and presenting different feature extraction methods relate with Devnagari script and proposed efficient and optimized extraction method with their comparative analysis. The accuracy of recognition system is majorly depending on feature extraction phase, types of features and size of features. For our research we are using hybrid efficient, faster and optimized feature vector which is combination of geometrical features, regional features, distance transform and gradient features. Feature vector length is 91. In addition to this, in existing cases, the time required for extracting the geometrical features is very high; however we are using Universe of discourse to speed up the retrieval. From practical analysis, accuracy of proposed feature vector set is improved as compared to existing feature vectors.

© 2015, IJARCSSE All Rights Reserved

A. Regional/Structural Features

These features are also called as structural features. Below four features are considered as regional properties of input image.

1) *Euler Number:* This feature is nothing but difference of number of holes and number of objects in input image.
2) *Regional Area*: This is nothing but ratio of number of white pixels in image to the total pixels.
3) *Eccentricity:* It is used to for the smallest ellipse that fits the essential of the image.
4) *Orientation:* The angle between the x-axis and the major axis of ellipse that has the similar second moments as the region. Forming final feature vector using these four features called ReF.

B. Statistical /Geometric Features

These features are also called as geometric features. There are total nine features extracted for each zone. Total numbers of zones to be considered are 9. Therefore 81 features are extracted for input segmented image. For statistical feature extraction, first the segmented handwritten input image is skeletonized. Then universe of discourse will be applied on that image. Then image is divided into 9 equal size zones. These are summarized as below,

- From each zone, starters, intersections, and minor starters are extracted and then these features are Stored in one vector.
- Line segments extraction is done from image and stored into one vector for each zone.
- Line type detection is performed from line segments such as horizontal, vertical, right diagonal and left diagonal etc.
- The total numbers of each line type are extracted.
- Normalized length of each line type is calculated.

All 81 features are Stored into vector GeF. In above algorithm Skeletonization and universe of discourse are used to reduce the features extraction time. The entire skeleton in that zone should be traversed to extract different line segments in a particular zone. Usually some specific pixels in the character skeleton were marked as starters, intersections and minor starters.

1) *Universe of Discourse:*
   Universe of discourse is defined as the shortest matrix that covers the whole character skeleton. As the features extracted from the character image contain the positions of various line segments in the character image. Therefore first universe of discourse is selected. Then the image is divided into windows of equal size, and the feature extraction is performed on individual windows.
2) *Zoning*
   It is defined as the ratio of number of black pixels in each zone to the total number of black pixels in the whole image. Then the whole image is divided into windows of equal size and feature extraction is applied to each individual zone rather than the whole image. In our work, the image was partitioned into 9 equal sized windows. The divided zone is then taken as one sub image and suitable features are extracted. This process is repeated for all the remaining zones.
3) *Character traversal*
   Character traversal starts after zoning by which line segments in each zone are extracted. First, the starters and intersections in a zone are identified and then populated in a list. Algorithm starts by considering the starter list. Once all the starters are processed, minor starters obtained along the course of traversal are processed. The positions of pixels in each of the line segments obtained during this process are stored. Once all the pixels in the image are visited, the algorithm stops.
4) *Starters*
   Starters are usually those pixels which are having one neighbour in the character skeleton. All the starters in the particular zone are finalized and then populated in a list. After that character traversal is performed.
5) *Intersections*
   The definition for intersections is a pixel having more than one neighbour. A new property called true neighbours is defined for each pixel. The neighbouring pixels are categorized into as, direct pixels and diagonal pixels. All pixels in the neighbourhood of the pixel under consideration in the horizontal and vertical directions are considered as direct pixels. The remaining pixels in the neighbourhood which are in a diagonal direction to the pixel under consideration are called as Diagonal pixels.
6) *Minor starters*
   Minor starters are formed when pixel under consideration have more than two neighbours.

C. Gradient Feature Extraction

The gradient feature decomposition was originally proposed for online character recognition. Conventionally, the gradient is calculated on each pixel of the image. The gradient is a two-variable function where the image intensity function is at each image point. The gradient vector points in the direction of largest possible intensity increase at each image point. The gradient features of an input image are nothing but the measure of direction and magnitude of major change in image intensity at every pixel. The gradient direction at any pixel (x, y) gives the direction of greatest change in image intensity. The below equation is used to calculate the gradient of features.

© 2015, IJARCSSE All Rights Reserved

$$\Theta(x,y) = \tan^{-1}\frac{G_y(x,y)}{G_x(x,y)}$$

$$G_x(x,y) = \frac{\partial I(x,y)}{\partial x} \quad , G_y(x,y) = \frac{\partial I(x,y)}{\partial y}$$

The angle Θ is measured with x-axis (horizontal axis). The direction of the edge at a pixel (x, y) is perpendicular to the gradient vector at that point. Where Gx and G y are gradient components along x-axis and y-axis and are obtained at any pixel (x, y) by convolving the given image with 3×3 windows. The size of direction and magnitude is same as input image size. Hence to make process faster, we further applied 2D mean and 2D standard deviation to extract four gradient features. The gradient measures the magnitude and direction of the greatest variation in intensity of each pixel.

D. Distance Transform

This process first converting binary input image into gray level distance map. A distance transform assigns to each white pixel (background) of a binary image a value equal to its distance to the nearest black pixels (foreground) according to a defined metric. Distance transform gives us new image array which is same size of original binary image. Next we applied 2D mean and 2D standard deviation; this gives us two distance transform features. To facilitate the extraction of direction features, the following steps were required to prepare the character pattern:

1) Starting point and intersection point location
2) Distinguish individual line segments
3) Labeling line segment information
4) Line type normalization

After extracting all this features, we combining them into final array of total 91 vectors: Statistical features (81) + Regional Features (4) + Gradient Features (4) + Distance Transform (2).

E. Formation of Feature Vector Size

For better accuracy, it is consider that more features are required with less time of feature extraction. Each feature is used to form a feature vector hence if we use a combination of features then it will help us to derive the feature vectors with more elements which are helpful to increase the efficiency of recognition.

IV. PROPOSED ALGORITHM OF FEATURE EXTRACTION

We are following holistic approach in which character and words are used as whole single unit and recognized using features extracted from it. In this paper we are using different types of features and combined them to form one final 91 features vector for recognition. All features used are listed below.

PROPOSED FEATURE EXTRACTION ALGORITHM

Input: Segmented Handwritten Image
Step 1: Extract Statistical/Geometric Features from Input and form feature vector GeF
Step 2: Extract Regional/Structural Features from Input and form feature vector ReF
Step 3: Extract Gradient Features: Gradient and Direction
Step 3.1: Apply mean and standard deviation on gradient
Step 3.2: Apply mean and standard deviation on direction
Step 3.3: Form final 4 features gradient vector called GrF
Step 4: Extract Distance Transform from Input.
Step 4.1: Apply mean and standard deviation on distance transform
Step 4.2: Form final 2 features distance transform vector called DiF.
Step 5: Combine GeF, ReF, GrF and DiF to form 91 feature vectors called CHF.
Output: Feature Vector CHF.

V. RESULTS AND DISCUSSION

In this section we are presenting results of various feature extractors. Table 5.1 shows the result of regional features such as Euler number, regional area, eccentricity, Orientation. Table 5.2 shows results of gradient feature extractor like gradient Magnitude & gradient direction. Table 5.3 shows the output results of distance transform feature extractor. Table 5.4 shows the output results of computation of elapsed time.

TABLE 5.1 RESULTS OF REGIONAL /STRUCTURAL FEATURE EXTRACTION (REF)

Sr. No.	Input Image	Euler Number	Regional Area	Eccentricity	Orientation
1	अमर	3	4593	0.8038	3.16
2	मुलगा	9	4877	0.7221	-63.11
3	नदी	1	4078	0.7361	14.12
4	दोन	7	4011	0.9950	90

© 2015, IJARCSSE All Rights Reserved

TABLE 5.2 RESULTS OF GRADIENT FEATURE EXTRACTION (GRF)

Sr. No.	Input Image	Gradient Magnitude (Gmag)	Gradient Direction (Gdir)
1	अमर	0.0513	0.2252
2	मुलगा	0.0472	0.2633
3	नदी	0.0494	0.246
4	दोंड	0.0654	0.2153

TABLE 5.3 DISTANCE TRANSFORM FEATURE EXTRACTION (DIF)

Sr. No.	Input Image	Dt1	Dt2
1	अमर	95.107788	4.578
2	मुलगा	92.890	10.63
3	नदी	100.70	5.211
4	दोंड	103.601	4.0112

TABLE 5.4 COMPUTATION OF ELAPSED TIME

Sr. No.	Input Image	Elapsed time
1	अमर	0.939231
2	मुलगा	1.665607
3	नदी	0.926243
4	दोंड	1.005036

VI. RESULTS AND DISCUSSION

In this section we are discussing the practical environment, scenarios, performance metrics used etc. In this section we will present the current results and comparison with existing method. Below figure 6.1 is showing the outputs of handwritten image recognition [Feature extraction phase].

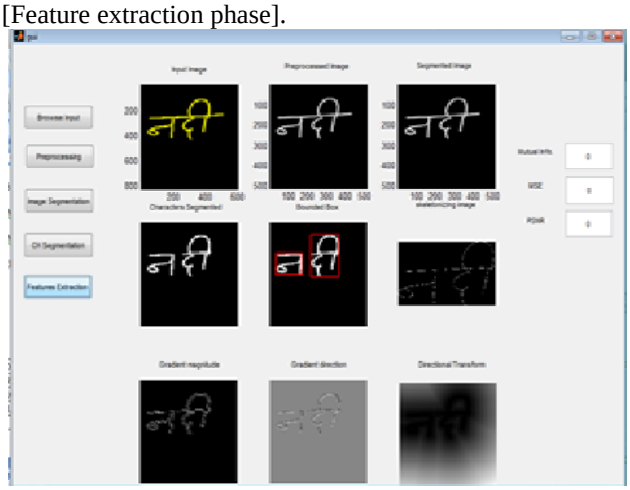

Figure 6.1: Results of segmentation and feature extraction for various images

This figure is showing number of features extracted based as per we proposed in above sections.

Below table 6.2 showing the comparative study of recognition accuracy by assuming that we are using KNN classifier for classification. In below table, GeF-Geometric/Statistical Features, ReF means regional features, GrF means gradient features, DiF means distance transform features, VHF means recently presented hybrid feature vector and CHF means proposed optimized, faster feature vector.

TABLE 6.2: RECOGNITION ACCURACY ON DIFFERENT DEVANAGARI SCRIPT DATASETS

Devanagari Dataset	GeF	ReF	GrF	DiF	VHF	CHF
Dataset1	96.56	91	94	92	94.2	96.8
Dataset2	96.67	94.76	94	93.2	97.02	97.12
Dataset2	91.02	89	90	88.3	90.88	92.1

Above table is showing the tentative comparative results for different kinds of features vectors against proposed feature vector. Proposed feature vector is showing better recognition accuracy as compared to existing methods.

© 2015, IJARCSSE All Rights Reserved

VII. CONCLUSION AND FUTURE WORK

In this paper we discussed about various feature extraction method related with handwritten word recognizer for Devnagari script. As our research area is online handwritten word recognition on Devanagari script, we have proposed an efficient, faster algorithm to extract the feature. The proposed algorithm is the combination of statistical features, structural features, gradient features, and distance transform features etc. To minimize the feature extraction time specially required for geometric features we have applied universe of discourse & skeletonizing of input image. The scope of this paper was limited to investigation and presenting new feature extraction algorithm. For future work, we suggest to work for minimization of extraction time required so that overall process will become faster with maximum accuracy.

REFERNCES

[1] Satish Kumar , "Performance Comparison of Features on Devanagari Hand-printed Dataset", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009, © 2009 ACADEMY PUBLISHER

[2] J.Pradeep, E.Srinivasan, "Diagonal Feature Extraction Based Handwritten Character System Using Neural Network", IJCAs (0975 – 8887) Volume 8– No.9, October 2010,PP.17-22

[3] Brijmohan Singh, Ankush Mittal, "An Evaluation of Different Feature Extractors and Classifiers for Offline Handwritten Devnagari Character Recognition", Journal of Pattern Recognition Research 2 (2011) 269-277

[4] Vandita Singh, Bhupendra Kumar, Tushar Patnaik , "Feature Extraction Techniques for Handwritten Text in Various Scripts: a Survey", IJSCE, ISSN: 2231-2307, Volume-3, Issue-1, March 2013

[5] Gita Sinha, Mrs. Rajneesh Rani, Prof. Renu Dhir, "Handwritten Devanagari Numeral Recognition Using Zonal Based Feature Extraction Method and SVM Classifier",© 2012, IJARCSSE, Volume 2, Issue 6, June 2012 ISSN: 2277 128X

[6] Pratibha Singh "Reliable Devanagri Handwritten Numeral Recognition using Multiple Classifier and Flexible Zoning Approach", I.J. Image, Graphics and Signal Processing, 2014, 9, 61-68, DOI: 10.5815/ijigsp.2014.09.08

[7] Munish Kumar, R. K. Sharma1 and Manish Kumar Jindal2 "A Novel Feature Extraction Technique for Offline Handwritten Gurmukhi Character Recognition", IETE JOURNAL OF RESEARCH | VOL 59 | ISSUE 6 | NOV-DEC 2013

© 2015, IJARCSSE All Rights Reserved