International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Recognition of Word using different Feature Extraction Methods

Kavita Babalad[1], Laxmidevi Noolvi[2], Dr. Jharna Majumdar[3]

[1]Department of CSE (PG), Nitte Meenakshi Institute and Technology, Bangalore, India

[2]Department of CSE (PG), Assistant Professor, Nitte Meenakshi Institute and Technology Bangalore, India

[3]Dean R&D, Prof and Head CSE (PG), Nitte Meenakshi Institute and Technology, Bangalore, India

Abstract: ***Word Recognition is one of the most interesting and challenging research areas in the field of image processing. In this paper, we are presenting a printed Word Recognition method, where we segment each character from the word. For segmented character each invariant features like line segments, different types of line segments, normalized horizontal and vertical line, crossing, distances and moment of inertia are been extracted. The clustering methods are used for recognition. The recognized word is spelled out by the system.***

Keywords: Pre-processing, Segmentation, Feature Extraction, Recognition and Clustering method.

1. Introduction

Word Recognition is an art of detecting segmenting and identifying characters from image [1]. It is a technique that provides full alphanumeric recognition printed characters at electronic speed by simply scanning the form or document. Word recognition is one of the most fascinating and challenging areas of pattern recognition with various practical application potentials. It can contribute immensely to the advancement of an automation process and can improve the interface between man and machine in many applications. The character recognition system helps in making the communication between a human and a computer easy.

An image processing technology called word recognition is used to identify the character in a word. There exist many feature extraction methods which have their own advantages and disadvantages over other methods. There are several important criteria of feature extraction methods required to be considered for higher recognition rate. Firstly, an effective feature needs to be invariant with respect to character by various writing styles of different individuals. It also needs to represent the raw image data of character through a reduced set of information. Where feature are to be extracted for each character and the characters are grouped and word is recognized. Extracting scale invariant features and different clustering methods are applied for Recognition. Word recognition field, emphasizing the methodologies required for the increasing needs in newly emerging areas, such as development of electronic libraries, multimedia databases, data entry, text entry and sign board identification.

2. Proposed Method

This paper discusses various methodologies for recognition of printed characters. The Input word image is pre processed and then the word is segmented into individual characters. Extracting the scale invariant features for proper identification and clustering techniques are used for Recognition of word and audio will be played for recognized word.

The below flowchart explains the flow of the word recognition. The details of flowchart are explained in the further sections.

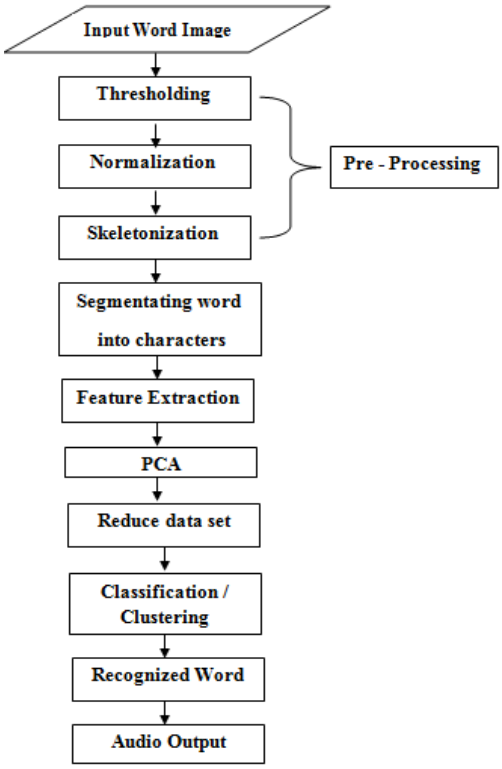


Figure 1: Overview for word recognition

1) Pre- Processing

Pre-Processing perform some operations such as thresholding and Skeletonization

- **Thresholding** is an important process as the results of the following recognition is totally dependent on the quality of the image, where a grayscale image is converted into binary image. It is done in order to identify the objects of

interest from the image. It separates the foreground pixels from the background pixels.

- **Skeletonization** is a morphological operation in which a single pixel wide representation of an image is obtained without changing its connectivity. The purpose of thinning is to reduce the image components so that they contain only essential information. The commonly used thinning algorithms are the Classical Hilditch algorithm, Zhang-Suen algorithm, Stentiford thinning algorithm. A survey of thinning methodologies is presented in [8]. Skeletonization process by which a one-pixel width representation (or the skeleton) of an object is obtained, by preserving the connectedness of the object and its end points [5]. Hence, this process can be seen as a conditional deletion of boundary pixels. Skeletonization image is shown in the Fig. 3

P9	P2	P3
P8	P1	P4
P7	P6	P5

Figure 2: Given 3×3 window showing the 8-neighborhood of pixel P1

Figure 3: Input image Skeletonization Image

2) Segmentation

Segmentation is one of the most important phases of word recognition system. By applying good segmentation techniques, we can increase the performance of recognizing the printed characters. Segmentation subdivides an image into its constituent regions or objects. Basically in segmentation, we try to extract basic constituent of the script, which are certainly characters. This is needed because our classifier recognizes these characters only. Segmenting the word into individual characters is done by vertically and horizontal scanning. Vertical scanning is done before horizontal scanning because if the skewness is present in the input image then its effect will be minimized [6].

Figure 4: Vertical Scanning

Figure 5: Horizontal Scanning

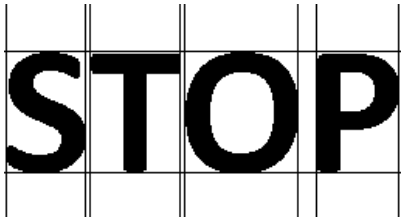


Figure 6: Vertical and Horizontal Scanning

- **Vertical Scanning**
  Scan the image from left or right until we get the continuous transition from white to black pixels. Once the transitions from white to black pixel occur draw the line and scan till the continuous white pixel occurs. After we get the continuous transition from black to white at the right draw the line. This process continuous untill we get the white pixels.
- **Horizontal Scanning**
  Scan the image from top until we get the continuous transition from black to white pixels. Once the transition from black to white pixels occurs draw the line and scan till the continuous white pixel occurs. After we get the continuous transition from black to white at the bottom draw the line at the bottom.
- **Normalization**, where each segmented character is normalized to fit within suitable matrix like 32x32 or 64x64 so that all characters have same data size.

3. Feature Extraction Methods

Features extraction is the important method which is used to extract the most relevant features which is further used to classify for the recognition process. In our research work, we have used the following listed features.

3.1 Intersections

An intersection is based on number of true neighbors for a particular pixel. Scan the image from top to bottom using 3x3 masks. Find the image pixel having three or four neighbor pixel and the number of transition (1 to 0 or vice versa).

Figure 7: Intersection point

3.2 Open ends

The image pixel having only one neighboring pixel is the open end. As shown in the below Fig 8 the number of open ends is represented [2].

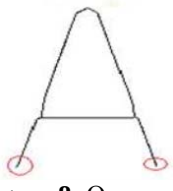
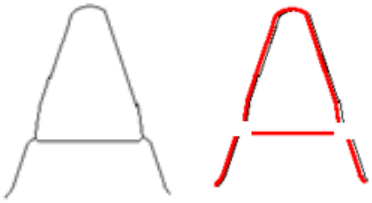


Figure 8: Open ends

3.3 Line Segment

To extract the number of line segments the thinned character image is given as the input [2]. Line Segment is calculated on the basis of number of intersection points and the open ends. Take 3x3 window at the intersection points. Here we are tracing the line until we get continuous white pixel. When the white pixel occurs, that is the end point of the line.



(a) (b)

Figure 9: (a) Skeletonization image (b) Number of line segments

3.3.1 Different Line Segments

After line segments have been extracted from the image they have classified into the following line types [2].

1. Horizontal Line
2. Vertical Line
3. Right diagonal line
4. Left diagonal line

For this, a direction vector is extracted from each line segment which will help in determining each line type. For this, a convention is required to define the position of a neighboring pixel with respect to the center pixel of the 3x3 matrix under consideration. The naming convention is as follows.

4	5	6
3	C	7
2	1	8

Figure 10: Naming Convention

In the matrix given, 'C' represents the center pixel. The neighboring pixels are numbered in a clockwise manner starting from pixel below the central pixel. To extract direction vector from a line segment, the algorithm scan through the entire pixels in the line segments in the order they forms the line segment [2].

The line segment marked in the image was obtained before applying the direction rules [7] explained last. But after applying the direction rules explained here, the two line types will be differentiated. If a new line segment is detected, then the direction vector is broken down into two different vectors at that point. Now the following rules are defined for classifying each direction vector.

1) If maximum occurring direction type is 2 or 6, then line type is right diagonal.
2) If maximum occurring direction type is 4 or 8, then line type is left diagonal.
3) If maximum occurring direction type is 1 or 5, then line type is vertical.
4) If maximum occurring direction type is 3 or 7, then line type is horizontal.

3.3.2 Normalized Length for different line segments

The number of any particular line type is normalized using the following method [2],

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)

3.4 Crossings

Crossings are one of the popular statistical features for recognizing printed characters [3]. It is defined as number of transition from background to foreground or foreground to background along a straight line throughout the image[12]. In other word it counts the number of stroke on a line from one side to another side thought the image. In this experiments crossing is computed for every column and row to construct the feature vector of the image [4]. Unlike other features this feature is not influenced by the width of strokes and can be computed without skeletonising the image. For a Single column and row is as shown in below fig 11.

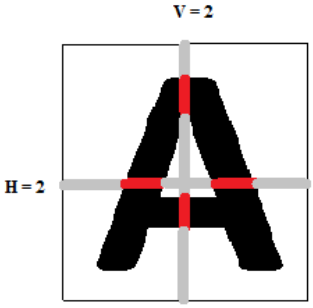


Figure 11: Crossings

3.5 Distances

The distances of the first image pixel detected from the upper and lower boundaries of the image, along vertical lines and from the left and right boundaries along horizontal lines [12].

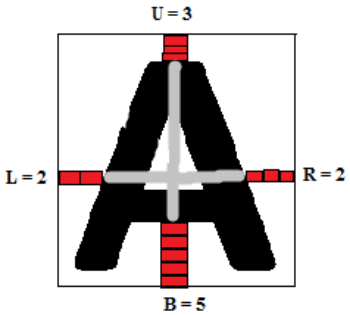


Figure 12: Distances

3.6 Moment of inertia

Regional moment of inertia [9] captures spatial information about the weight distribution of the character at different positions along its vertical axis. Because characters maintain constant orientation and are generally symmetric about their vertical axis, Ixx quantities for six different regions are used as a descriptor. It is scale invariant. Fig.13 represents Regional moment of inertia.

$$I_{xx} = 1/N \sum_{xi \in R} (X_i - X_{centrod})^2$$

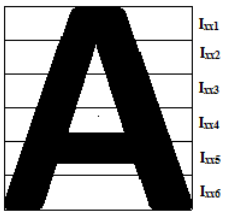


Figure 13: Moment of Inertia

3.7 Aspect Ratio(Slimness)

The aspect ratio is a ratio between the maximum length Dmax and the minimum length Dmin of the minimum bounding rectangle (MBR).

$$\text{Aspect ratio} = \frac{D_{max}}{D_{min}}$$

3.8 Rectangularity

Rectangularity is the measure of how the character approaches to rectangle or it can be defined as the similarity between character and rectangle. To calculate the rectangularity first step is to find the area of the image find the height and width of the image, using which the area of the minimum bounding rectangle can be found. Rectangularity is given by the ratio of the area of the image to the area of the minimum bounding rectangle

Rectangularity can be calculated

$$\text{Rectangularity} = \frac{A_S}{A_R}$$

4. Word Recognition Phase

4.1 Feature data set reduction (Principle Component Analysis)

The main principle of PCA is to reduce the dimensionality of the data while retaining as much as possible of the variation present in the original dataset [13]. Intuitively, Principal components analysis is a method of extracting information from a higher dimensional data by projecting it to a lower dimension. This method generates a new set of variables, called principal components. All the principal components are orthogonal to each other, so there is no redundant information. The objective of PCA is to perform dimensionality reduction while preserving as much of the randomness in the high-dimensional space as possible but the limitation with PCA is it depends on scaling of variables and it is not always easy to interpret principal components.

The principal components as a whole form an orthogonal basis for the space of the data. Mathematically, PCA transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate. Each coordinate is called a principal component. PCA allows us to compute a linear transformation that maps data from a high dimensional space to a lower dimensional sub-space.

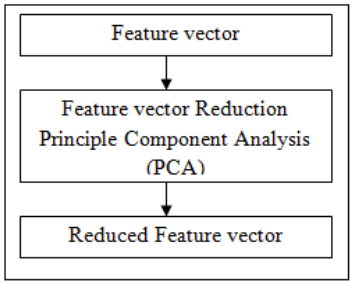


Figure 14: PCA reduced data feature vector

4.2 Clustering Method

The goal of the clustering analysis is to divide a given set of data or objects into a cluster, which represents subsets or a group [10]. The partition should have two properties like homogeneity inside clusters: the data, which belongs to one cluster, should be as similar as possible and heterogeneity between the clusters: the data, which belongs to different clusters, should be as different as possible.

The membership functions don't reflect the actual data distribution in the input and the output spaces. They may not be suitable for fuzzy pattern recognition [14]. To build membership functions from the data available, a clustering technique may be used to partition the data, and then produce membership functions from the resultant clusters. Thus, the characters with similar features are in one cluster. Thus, in recognition process, the cluster is identified first and then the actual character.

In this paper we have explained three kinds of clustering methods they are K-means and Fuzzy C-Means.

K-means is a simple unsupervised learning method which can be used for data grouping or classification when the number of the clusters is known [10], where k stands for number of cluster. Thus, this method works for a fixed set of characters. Given a set of initial clusters, assign each point to one of them, and then each cluster centre is replaced by the mean point on the respective cluster. These two simple steps are repeated until convergence. A point is assigned to the cluster which is close in Euclidean distance to the point. Although K-means has the great advantage of being easy to implement, but has two major drawbacks. First, it can be really slow since in each step the distance between each point to each cluster has to be calculated, which can be really expensive in the presence of a large dataset. Second, this method is really sensitive to the provided initial clusters, however, in recent years, this problem has been addressed with some degree of success.

Fuzzy C-Means algorithm [3] works by assigning membership to each data point corresponding to each cluster centre on the basis of distance between the cluster centre and the data point [15]. More the data is near to the cluster centre more is its membership towards the particular cluster centre. Clearly, summation of membership of each data point should be equal to one.

4.3 Comparison between the K-means and Fuzzy C-Means algorithm

Table 1: Comparison between K-means and Fuzzy C-Means

K-means	Fuzzy C-Means
Centre of the cluster is the mean value of the objects in the cluster.	Centre of the cluster is based on the fuzzy membership Function.
Values of characteristic function range from 0 or 1.	Values of characteristic Function range from 0 to 1.
Elements may belong to only one cluster	Elements may belong to more than one cluster
The K-means requires only number of clusters, & randomly chosen cluster centres as inputs.	The Fuzzy C-means requires only number of clusters, & randomly chosen cluster centres as inputs
K-means clustering produces Fairly higher accuracy and requires less computation.	Fuzzy C means clustering produces close results to K-means clustering, yet it requires more computation time

4.4Testing or Recognition of word

After extracting features for each character, applying the PCA to get the reduced set of feature vector and applying k-means cluster. After this we have to go for recognition phase. It has following steps:

Step 1: Test Input word image pre processing.
Step 2: Test Input word image is segmented into characters and features are extracted for each individual characters.
Step 3: Features stored in database are retrieved for each character.
Step 4: Reduce the features of each character.
Step 5: Apply clustering or classification methods for each character.
Step 6: Compute Euclidean distances between cluster centers and test image features find minimum distance which indicates test image in that particular cluster.
Step 7: Go to that cluster and identify specified character.
Step 8: Repeat the steps 3 to 7 for all the characters in the word.
Step 9: Group the individual recognized characters to form a given word.
Step 10: Finally if the grouped characters are compared and matched with the given word then we can say the word is recognized properly.
Step 11: Finally the recognized word is spelled out by the system.

5. Results And Analysis

Comparison of recognition method is done based on the time taken for recognition rate of words. The experiment of Printed upper case characters and words contains 26 characters and 137 words samples in database for testing. For each query numeral, items in the database are sorted in order of increasing distance. Results are taken for different scale invariant features.

Recognition rate (RR) is defined as :

$$RR = \frac{Number\ of\ correctly\ recognized\ words}{Total\ number\ of\ testing\ words} \times 100$$

	No. of words	*No of words correctly detected*	*No of words incorrected*	*Recognition rate*
100% size words	137	122	15	89.05%
80% size words	137	107	20	78%

6. Conclusion

In this Paper, we have presented different invariant feature extraction methods like line segments, different types of line segments, normalize line segments, Crossings, distances and moment of inertia which improves the accuracy of feature extraction and clustering methods. This paper also provides the recognition of word and displays the results in the form of dictionary and gives the recognized word in the form of audio, in which the system spells out the recognized word. However the different methods of feature extraction, clustering methods and giving audio are discussed here are very effective and useful for new researchers.

References

[1] **"A Literature Review on Hand Written Character Recognition"** Mansi shah and Gordhan B Jethava Department of computer science & engineering parul Institute of Technology, Gujarat, India and Information Technology Department parul Institute of Technology, Gujarat, India Indian Streams Research Journal Vol-3, ISSUE- 2, March 2013 ISSN: 2230-7850

[2] **"A Feature Extraction Technique Based on Character Geometry for Character Recognition"** Dinesh Dileep Department of Electronics and Communication Engineering, Amrita School of Engineering, Kollam, Kerala, INDIA

[3] **"Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data"** International Journal of Database Theory and Application Vol.6, No.6 (2013), Lu, Tinghuai Ma1, Changhong Yin2, Xiaoyu Xie2, Wei Tian1 and ShuiMing Zhong School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044.

[4] **"A Comparative analysis of Feature Extraction techniques for Handwritten Character Recognition"**Rajbala Tokas1,Aruna Bhadu2 1 M.Tech*(CS),Swami Keshwanand Institute Technology,Jaipur,Rajasthan,India,2 M.Tech*(SE) Govt. Engineering College Bikaner,Rajasthan,India. International Journal of Advanced Technology & Engineering Research (IJATER) ISSN No: 2250- 3536 Volume 2, Issue 4, July 2012

[5] **"A Fast Parallel Algorithm for Thinning Digital Patterns"** research contribution Image Processing and Computer Vision Robert M. Haralick Editor author T. Y. ZHANG and C. Y. SUEN

[6] **"Pixel Clustering Based Partitioning Technique for Character Recognition in Vehicle License plate"** Siddhartha Choubey Associate Professor (CSE) Shri Shankaracharya College of Engg. & Technology,Bhagwati Charan Patel Associate Professor (IT) Shri Shankaracharya College of Engg. & Technology, G.R.Sinha Professor & Head (IT), IEEE Member Shri Shankaracharya College of Engg. &

Technology, Abha Choubey Associate Professor (CSE)Kavita Thakur Reader (SOS in Electronics), IEEE Member Pt. Ravishankar Shukla University, Raipur, Chhattisgarh, India 011 3rd International Conference on Machine Learning and Computing (ICMLC 2011)

[7] **"A Novel Feature Extraction Technique for the Recognition of Segmented Handwritten Characters"** M. Blumenstein, B. Verma and H. Basli School of Information Technology, Griffith University-Gold Coast Campus, Australia Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003) 0-7695-1960-1/03 $17.00 © 2003 IEEE

[8] **"Handwritten Character recognition in malayalam scripts – A Review"** International Journal of Artificial Intelligence and Applications(IJAIA), Vol. 5, No. 1,January 2014

[9] **"Automatic Plant Leaf Classification for a Mobile Field Guide"**. David Knight, James Painte, Matthew Potter

[10] **"Survey of Methods for Character Recognition"** Suruchi G. Dedgaonkar, Anjali A. Chandavale, Ashok M. Sapkal" **International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 5, May 2012**

[11] **"A Fast and Robust Scheme for Recognition of Handwritten Devnagari Numerals"** XXXII national systems conference, nsc 2008, December 17-19, 2008 C. Vasantha Lakshmi, Ritu Jain, and C. Patvardhan

[12] **"A Comparative Analysis of Feature Extraction Techniques for Handwritten Character recognition"**International Journal of Advanced Technology & Engineering Research (IJATER) ISSN No: 2250-3536 Volume 2, Issue 4, July 2012 Rajbala Tokas,Aruna Bhadu M.Tech*(CS),Swami Keshwanand Institute of Technology,Jaipur,Rajasthan,India, M.Tech*(SE) Govt. Engineering College

[13] **"A tutorial on Principal Components Analysis"** Lindsay I Smith February 26, 2002.

[14] **"An efficient k-means clustering algorithm Khaled Alsabti"** Syracuse University Sanjay Ranka University of Florida Vineet Singh Hitachi America, Ltd.

[15] **"Evaluation of Fuzzy K-Means and K-Means Clustering Algorithms In Intrusion Detection Systems"** Farhad Soleimanian Gharehchopogh, Neda Jabbari, Zeinab Ghaffari Azar international journal of scientific & technology research volume 1, issue 11, december 2012 issn 2277-8616