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

A New Approach for Sentence Recognition Using Multiple Features

Shruthi Shri .D[1], Dr. Jharna Majumdar[2], Laxmidevi Noolvi[3]

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

[2]Dean R&D, Professor and Head CSE (PG), Nitte Meenakshi Institute of Technology, Bangalore, India

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

Abstract*: This paper presents the work towards recognition of sentence by using segmentation based on bounding box for character recognition. After segmentation various variant and invariant features are extracted like Correlation method, Shadow feature, horizontal and vertical profiling, Chain code and Projection methods which increase the recognition capability of a character. Classification techniques are used for improving the recognition of sentence. Dictionary search for each sentence is performed, after first character is recognized and matched. The recognized sentence is spelled out by the system.*

Keywords: Segmentation, Feature Extraction, classification, Character Recognition

1. Introduction

Nowadays, sentence recognition systems are used in many fields that have different applications in robotics,license_plates, businesscards, invoices, ID_cards, driver licenses. The character recognition started from the recognition of machine printed characters and then it was developed to the recognition of the machine printed words and sentences. Gradually, handwritten digit, character and word recognition were introduced into this domain. Several research works have been focusing towards evolving the newer techniques that would reduce the preprocessing time and to provide higher recognition accuracy.

Text objects occurring in sentence images contain much information related to characters. Therefore, the text information has to be extracted efficiently, by performing pre-processing on the input sentence image. After segmenting sentence into words, words to isolated characters the features of each characters is extracted. Feature Extraction plays an important role in identification of character hence, selection of good feature set is most important aspect of printed character pattern recognition so Correlation method[6], horizontal, vertical profiling, Shadow feature[5], chain code and projection methods increases the recognition capability. The method as provides the ease of implementation and recognition [4]. Here we use the concept of clustering in identification of characters in a sentence.

2. Proposed Method

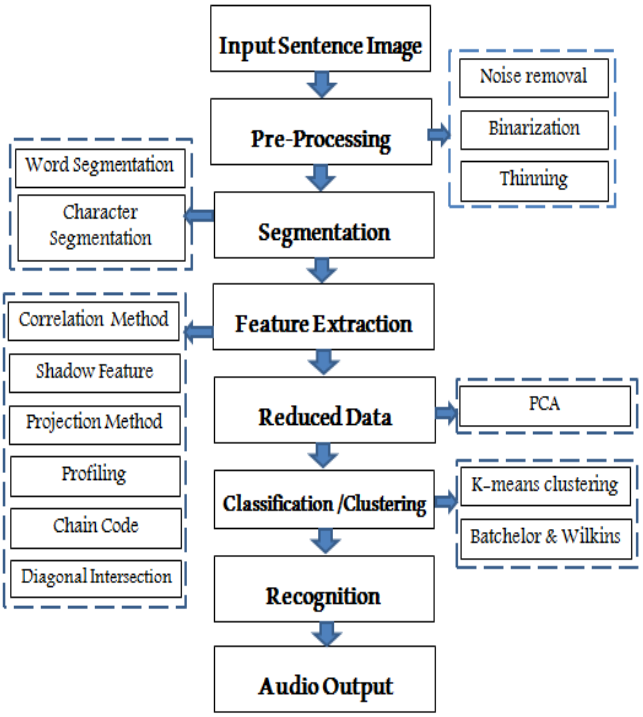


Figure 1: Flow Chart For Proposed Method

The proposed method consists of 6 main steps.

i. Preprocessing
ii. Segmentation
iii. Feature extraction
iv. Reduced Data
v. Clustering
vi. Recognition

A. Preprocessing

In the scanning process, some distorted images may be introduced due to light machine printing, quality of paper on which the character printed. Some Pre-processing steps are

performed for rectification of distorted images, improving the quality of images and size normalization. The goal of this pre-processing is to distinguish between image pixels that belong to text and those that belong to the background. Every image is first converted into a grayscale image to preserve memory and speed up further processing. After threshold is applied, the grayscale image converts into a binary image with only two possible values, black (0) and white (1).

B. Segmentation

Segmentation is an integral part of any text based recognition system. It assures efficiency of classification and recognition. Accuracy of character recognition heavily depends upon segmentation phase. Incorrect segmentation leads to incorrect recognition. Segmentation phase include line segmentation, character and word segmentation. It is important to obtain complete segmented character without any noise to ensure quality feature extraction.

The pre-processed image is considered for segmentation of words in a sentence. We use vertical and horizontal scanning to obtain the bounding box coordinates of each character. We can separate a word and character by space between them. The spacing between the word and the character will vary so drawing a bounding box for words and isolated characters in a sentence is quite challenging. Space between the characters 'WO', 'OW' 'AO','OV', 'VO' is perceived by the human eye we find some space between characters but a machine cannot identify the space and segment characters. Here in this paper we propose a concept of segmentation using bounding box to segment words and characters in a sentence as shown in Figure 2.

(a) **Bounding box Segmentation for word**

(b) Bounding box Segmentation for character

Figure 2: Bounding box Segmentation for word and character

C. Feature Extraction

It is finding the set of parameters that define the shape of character precisely and uniquely. Feature set plays one of the important role in recognition. Transforming the input data into the set of features is called Feature extraction. There are several feature extraction methods that are based on geometric, diagonal features and pixel based feature.

1. Correlation Function

The correlation coefficient is one of the popular metric used in the literature to provide comparison of two images. For our empirical knowledge it is fundamentally of co-varying things. We consider the skeletionized character image for correlation function. The character images size depends on font size of a character, define each character into different segments, i.e. of size 21*21. The two elementary shapes chosen are horizontal and vertical line. These entire row elementary shapes were located at central position the 21*21 window. Our feature extraction stage identifies and determines the type and location of the elementary shapes in the thinned character samples. To compute correlation, we need to define the correlation points in the skeletonized character image size.

Horizontal axis of normalized character image (m) …. (1)
Vertical axis of normalized character image (n) …. (2)

Combining the axes in equation 1 and 2, we obtain following correlation points around which we take the different segments of each character:

Now, we divide the character into different segments. Each segment is of size 21*21 with its centre at the point of 'mn' vector in equation(3). The graphical form of this whole procedure is shown in figure 3 which contains two elementary shapes on left side and example handwritten character on the right side. The example handwritten is divided into segment windows the centre of each segment window is indicated by a cross.

Equation for Correlation:

$$CF(m,n) = \frac{\sum_{x\in N}\sum_{y\in N} f'(x,y)g'(x,y)}{(\sum_{x\in N}\sum_{y\in N}(f'(x,y))^2(\sum_{x\in N}\sum_{y\in N}g'(x,y))^2)^{1/2}}$$

----- (1)

We find the normalized correlation value of different segments with elementary shapes by the following procedure:

Algorithm

Input: Thinned Character Image
Output: Correlation function based feature values
Step 1: Initially thinned character image is considered as input.
Step 2: The templates of size 30*30 is created i.e vertical line, horizontal line, positive slant line and negative slant line..
Step 3: Pixel by pixel move is made to match the template using correlation function based features from the thinned character.
Step 4: Correlation equation is calculated for every pixel move, if exact match for template is found then count is taken.
Step 5: Occurrence of template is counted for different templates.
Step 6: Counts are noted for different templates.

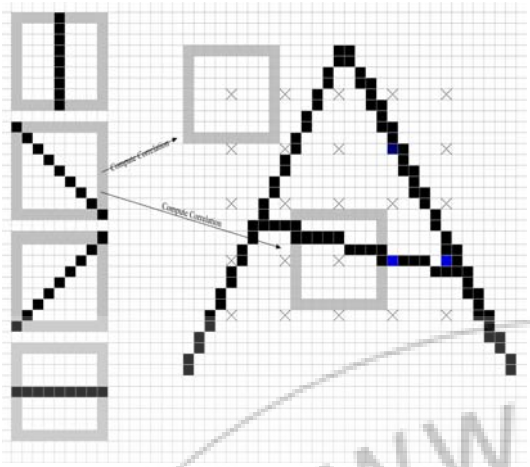


Figure 3: Normalized correlation of elementary shapes for a character

This gives a total horizontal template match of 54. An important property of these features, since they are based on variant in the character images.

2. Shadow Feature of a character

The shadow features for a character are computed globally and intersection features, line fitting features are computed by dividing the character image into 8 different segments. 16 shadow features are extracted from eight octants of the character image. The rectangular boundary enclosing the character image is divided into eight octants, for each octant shadow of character segment is computed on two perpendicular sides so a total of 16 shadow features are obtained shown in Fig. 4(a).

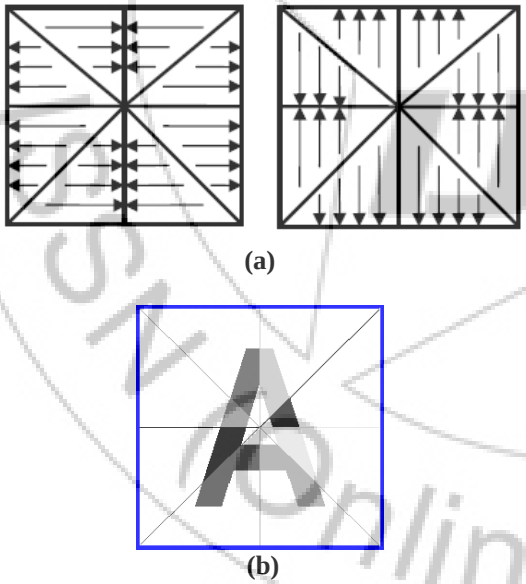


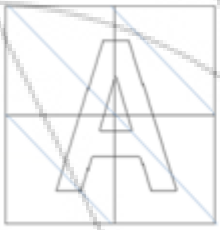
Figure 4: (a) Projections on sides (b) Character image is divided into eight octants, Shadow feature Extracted.

Shadow is basically the length of the projection on the sides as shown in figure. These features are computed on scaled image. Character recognition technique uses features obtained from Shadow, intersection and chain code histogram. After segmenting the character image into eight octants, each octant has the part of character black pixels i.e each octant's black pixel count is obtained. Then final step is repeated to all octants in the image and count is obtained. Shadow feature is extracted.

3. Projection Method

The projection method does the compression of the data through a projection. Black pixel counts are taken along parallel lines through the image area generating marginal distributions. The direction of projection can be horizontal axis, vertical axis and diagonal axis. Initially character is extracted with the boundary points. Character image is segmented into four quadrants such that each and every line of a character image is scanned and black pixel counts are taken along parallel lines. As shown in Fig 4.

The direction of projection can be horizontal axis, vertical axis and diagonal axis.



(a)

1st Quadrant diagonal black pixels	2
2nd Quadrant diagonal black pixels	0
3rd Quadrant diagonal black pixels	2
4th Quadrant diagonal black pixels	2

(b)

Figure 5: (a) Projection Method, 5 (b) Values of each quadrant for diagonal projection.

Here the feature of character boundary black pixel count for horizontal and vertical projection are equal. So, instead we consider the diagonal black points as shown Figure 5. As a feature of character.

4. Horizontal and Vertical Profiling

The four profiles of one connected component encompassed with a bound box are obtained by counting the white pixels in the four directions, rightward, leftward, upward and downward respectively until the black pixels are encountered. Calculating right and left profile gives the horizontal profiling value, upward and downward profile gives the vertical profiling values. These two values are scale invariant for any character size image.

Horizontal Profile = (Right view)/(Left view)
Vertical Profile = (Upside view)/(Downside view)

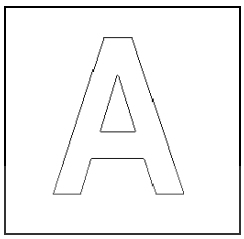
5. Chain Code Method

Chain Code, is a feature extraction method is investigated based on 4-neighborhood or 8–neighborhood methods. 8–neighborhood method has been implemented which allows generation of eight different codes for each character. These codes have been used as features of the character image.

Algorithm
Input: Character Boundary Image
Output: Normalized chain code values.

Step 1: Find out starting point which has non zero values and store it in array.
Step 2: Initialize 0-7 total eight directions.
Step 3: Travels all 8 neighbours.
Step 4: Find first nonzero value.

Step 5: Add it to the chain code list.
Step 6: Move to next position.
Step 7: Check whether we reach to first point or not if not then go to step 3.



0 1 1 2 2 2 2 2 1 2 1 2 1 2 2 1
2 2 1 2 1 2 1 2 1 2 1 0 1 5 4 5
6 5 6 5 6 5 6 5 6 6 5 6 6 5 6 5
6 5 6 6 6 6 6 5 5 4.....

Sample of Chain Code for Figure 5

Figure 5: Boundary of a character

Normalized chain code:

Table 1: Normalized chain code

0	1	2	3	4	5	6	7
2	12	15	0	2	12	15	0

D. Reduced Data

PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. PCA is that once we have found these patterns in the data, and we compress the data, ie. by reducing the number of dimensions, without much loss of information. this method of extracting information from a higher dimensional data by projecting it to a lower dimension. 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. Here higher dimension feature extracted data set is considered as input to reduce it to lower dimension. Purpose is to reduce the computation time and recognition rate

E. Clustering

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. Clustering is the process of identifying each character and assigning it the correct cluster group. K-means clustering is simple unsupervised learning method which can be used for data grouping or classification when the number of the clusters is known [8]. Where 'K' stands for number of clusters. 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, it has two big 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.

Batchelor and Wilkin's clustering method which can be used for data grouping or classification when the number of the clusters is unknown. Initial clusters are assigned internally i.e by supervised method. The cluster centers are calculated, steps are repeated until convergence. A point is assigned to the cluster which is close in Euclidean distance to the point. Then characters are clustered.

F. Recognition

This paper present an approach for sentence recognition, here each sentence and every character in a sentence is recognized by integrating feature extraction values, reduced data and clustering method.

The recognition of very first character in a sentence by using features and clustering method, the search i.e dictionary search is performed for a first character match and lists out the vocabulary of all sentences and words from the database. It compares for all characters in a sentence, the exact match is found in a database then a sentence is recognized.

3. Result and analysis

The proposed system analysed by calculating the recognition rate for characters, words and sentence images. Recognition rate (RR) calculated by :

$$RR = \frac{\text{Number of Recognised Sentence}}{\text{Total Number of Testing Sentences}} * 100$$

Table 2: Recognition rate for characters, words and sentences

Input Image	*Output analysis for proposed method*		*Recognition Rate*
	No. of Input Images	*No. of Recognized output*	
Characters	26 characters	26	99.73%
Words	40 words	40	95%
Sentences	60 sentences	58	95%

In proposed system, comparison of recognition method is done based on the time taken for recognition of sentence. Table (3) shows the comparison of the clustering of with and without PCA.

Table 3, illustrates different clustering methods, with including and excluding PCA. From observation of recognition rate in table K-means with PCA integrated is efficient compared to other clustering method.

Table 3: Recognition rate for sample sentences

Sl no.	*Clustering Method*	*Input Sentence*	*Time to Recognition*
1.	K-mean	Go Straight	0.44999
		Follow Right	0.42800
		U Turn	0.17609
		Go Slow	0.21700
2.	Bachelors	Go Straight	0.30400
		Follow Right	0.33909
		U Turn	0.15907
		Go Slow	0.18900
3.	K-mean with PCA	Go Straight	0.02987
		Follow Right	0.04508
		U Turn	0.01680

		Go Slow	0.02086
4.	Bachelors with PCA	Go Straight	0.38360
		Follow Right	0.38630
		U Turn	0.15200
		Go Slow	0.29110

4. Conclusion

The experimental results show that the algorithm successfully recognises sentence by implementing multiple features and clustering methods. The algorithm worked well for printed characters. The achieved time to identify is remarkable and will make the technique to apply in concerned application. It is difficult to achieve a robust algorithm that would be good in all segments and all cases of sentence recognition without fault. This research on sentence recognition improves recognition performance without faults and accuracy of recognition system by multiple feature extraction and clustering methods.

References

[1] "**Pixel Clustering Based Partitioning Technique for Character Recognition in Vehicle License Plate**" Siddhartha Choubey, G.R.Sinha IEEE Member, Bhagwati Charan Patel, Abha Choubey, Kavita Thakur IEEE member. 2011 3rd International Conference on Machine Learning and Computing (ICMLC 2011).

[2] "**Machine Printed Character Segmentation Method using Side Profiles.**" Min-Chul Jung, Yong-Chul Shin and Sargur N. Srihari State University of New York at Buffalo Buffalo, New wwYork 14260 U.S.A. 1999 IEEE.

[3] "**Character segmentation using side view feature in machine-printed optical character recognition**" Minchul jung, Department of Computer Science and engineering, sangmyung university, october 2010.

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

[5] **"Combining Multiple Feature Extraction Techniques for Handwritten Devanagari Character Recognition"** Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, Dipak Kumar Basu, Mahantapas Kundu. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-2, May 2013.

[6] **"Handwritten Character Recognition Using Multiclass SVM Classification with Hybrid Feature Extraction"** Muhammad Naeem Ayyaz, Imran Javed and Waqar Mahmood., Pak. J. Engg. & Appl. Sci. Vol. 10, Jan., 2012 (p. 57-67).

[7] **"Character Energy and Link Energy-Based Text Extraction in Scene Images"** Jing Zhang and Rangachar Kasturi, R. Kimmel, R. Klette, and A. Sugimoto (Eds.): ACCV 2010, Part II, LNCS 6493, pp. 308–320, 2011.

[8] "**An efficient k-means clustering algorithm**" Alsabti, Khaled; Ranka, Sanjay; and Singh, Vineet (1997). Electrical Engineering and Computer Science. Paper 43.

[9] **"A Two-Stage Segmentation Technique For Printing Kannada Text**" R Sanjeev kunte, Sudhakar Samuel R D, S J College Of Engineering, GVIP Special Issue on Image sampling and segmentation, March, 2006.

[10] **"Character Recognition Based On Region Pixel Concentration For License Plate Identification"** Kresimir romic, Irena Galic, Alfonzo Baumgatner. Tehnicki vjesnik 19, 2(2012), 321-325.

[11] "Evaluation of Classification and Feature Extraction Techniques for Simple Mathematical Equations"

[12] Sanjay S. Gharde, Baviskar Pallavi V, K. P. Adhiya, Department of Computer Engineering, SSBT, s College of Engineering & Technology, Jalgaon, Maharashtra State, INDIA. International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 1– No.5, February 2012 – www.ijais.org.

[13] **"Handwritten Character Recognition System using Chain code and Correlation Coefficient"**Ravi Sheth , N C Chauhan, Mahesh M Goyani, Kinjal A Mehta. International Conference on Recent Trends in Information Technology and Computer Science (IRCTITCS) 2011 Proceedings published in International Journal of Computer Applications® (IJCA).

[14] **"A simplified Method for handwritten character recognition from document image",** Mohammad Imrul Jubair, Prianka Banik, international Journal of Computer Applications (0975-8887) Volume 51- No, 14, August 2012.

[15] **"Segmentation Techniques For Image Analysis: A Review", Jay Acharya, Sohil Gadhiya, Kapil Raviya, i**international Journal of Computer Science and Management Research Vol. 2 Issue 1 January 2013, **ISSN 2278-733X.**