International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 10, October-2015

A Survey on Machine Printed Gurmukhi Text Recognition

Triptinder Pal Kaur
Department of Computer Science and Engineering
GZS PTU Campus,
Bathinda (Punjab)

Dr. Naresh Garg
Department of Computer Science and Engineering
GZS PTU Campus,
Bathinda (Punjab)

Abstract **- Creating a paperless environment is requisite due to the presence of extensive applications of computers and multimedia techniques and it was comply with the evolution of optical character recognition .Character recognition may resolve numerous complex problems and furthermore make the human job effortless and faster .From the last few decades it has been a growing trend amid the world wide researchers to recognize characters from various scripts .This subject has attracted numerous researchers since it provides a mean for automatic processing of substantial amount of data .Machine printed character recognition has been a laborious research area in the field of pattern recognition .There is a rich literature accessible on recognition of machine printed words for Indian and non-Indian scripts such as English, Chienese, Arabic, Devnagri etc. ,apart from restricted work is accessible on the recognition of machine printed Gurmukhi words . Different recognition models have been put forth in recent years and different research groups are working on the recognition of Gurmukhi words .Feature Extraction and classification are the crucial phases of the character recognition process that influence the overall accuracy of the recognition process. This paper will provide the overview of different techniques accustomed for feature extraction and classification of Gurmukhi scripts by different researchers and the conclusion obtained by them.**

Keywords- Optical Character Recognition, Support Vector Machine, Artificial Neural Network.

I.INTRODUCTION

Earlier the keyboard entered data and the scanned documents were manually processed by human beings which was a time consuming process ,data acquisition was also slow and requisite human intrusion increased the chances of errors and all this was comply with the evolution of optical character recognition. OCR started from the recognition of machine printed digits and characters and then it was developed for recognition of machine printed words and then gradually recognition of characters and text from the images were introduced into this domain. Various commercial OCR systems for different Latin languages are accessible on our personal computers. Computer systems armed with OCR to enhance the speed of performing operations, reducing data entry errors and also reduce the storage space. Fast retrieval, flexibility, accuracy and speed are the three main features that makes a good OCR. OCR is a challenging area in the field of image processing. Basic steps performed during the process of character recognition are:-summarized as:-

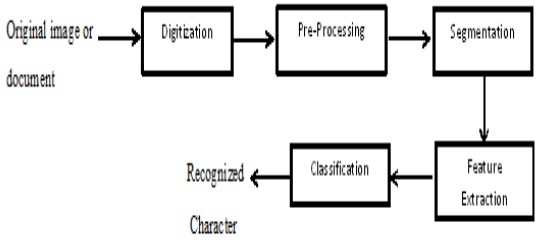

Fig. 1. Basic Steps for Recognition

A. Pre-processing

Pre-Processing phase embrace several operations that are applied successively on the input image for efficient character recognition .Some of the operations performed are:-

1) *Binarization:*-In this we transform the gray scale image to a binary image (0 and 1pixel values).

2) *Noise Removal:* - Occasionally the binary image provided as the input to the system might contain many imperfections due to presence of noise. So, noise removal operations remove unwanted bit patterns that have no significance in the output.

3) *Normalization:* - Input images may be of varying size. Normalization is basically performed to manage the consistent size of the character.

Due to presence of some imperfections in the image morphological operations can also be applied on the binary image. Some of the fundamental morphological operations are:-

4) *Skeletonization :* -It reduces the width of the object to a single pixel .It retains the significant information about the object besides It eliminates the irregularities and make the recognition process effortless.

5) *Thinning*: -It is another process analogous to skeltonization that reduces the width of the binary object to a single pixel.

6) *Erosion and Dilation*: - Erosion usually eliminates the small details from the image i.e. it just removes some pixels from the boundary of the image. Dilation is the reverse of erosion, it just add some pixels on the boundary of the image.

After preprocessing now the image is ready to fed to the next phase.

B. Segmentation

Segmentation phase is an important phase of OCR because improper segmentation can lead to misrecognition. It have a large influence on the general precision of the recognition process, because improper segmentation can lead to misrecognition. During recognition of scripts segmentation basically incorporates line, word and character segmentation.

C. Feature Extraction

In this procedure we select a set of specific features by which we can uniquely distinguish each character segmented. There are different feature extraction methods yet determination of a suitable strategy is one of the critical task which influences the recognition rate of the framework.

D. Classification

The features we extracted in the preceding stage are used to classify each segmented character according to the pre-defined rules. The decision rules are made such that there is less probability of misrecognition. There are different types of classification methods and even there are multiple classifiers that can be applied such as Support Vector Machine, Artificial Neural Network. Combination of such classifier's can also be applied for the classification.

II. CHARACTERISTICS OF GURMUKHI SCRIPT

Gurmukhi script is basically used for Punjabi language and is the world's 14th most widely spoken language. . The word Gurmukhi is the compound form of Guru and Mukh, which means sayings that came from the mouth of Guru. It is widely spoken in northern India ,Canada ,Pakistan ,England and many other countries worldwide .It is the first official script which has been adopted by Punjab state and the second most widely spoken language in many other states of north India .Some of the characteristics of Gurmukhi script are:-.

- It is always written from left to right.
- Gurmukhi script is cursive.
- Gurmukhi script consist of 9 vowels,3 semi vowels(sound modifier),3 half characters that lie at the feet of consonants as shown in figure 2 and figure 3.
- Most of the characters have a horizontal line at the upper part and most of the characters of the word are connected to this line called the headline and therefore there is no vertical inter character gap between the letters of a word.
- A Gurmukhi word can be partioned in to three horizontal zones namely upper ,middle and lower zone.

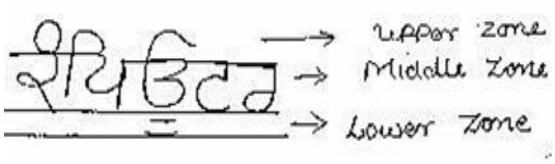

Fig 2. Horizontal Zones

- There is topologically similar pair of characters.

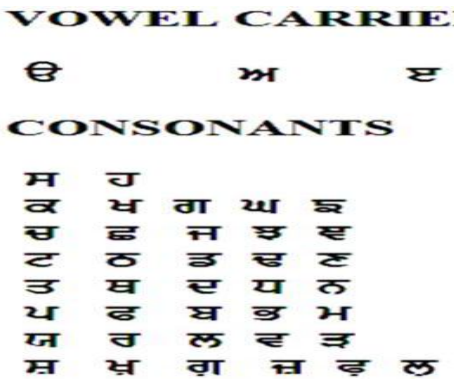

Fig. 3. Gurmukhi Characters

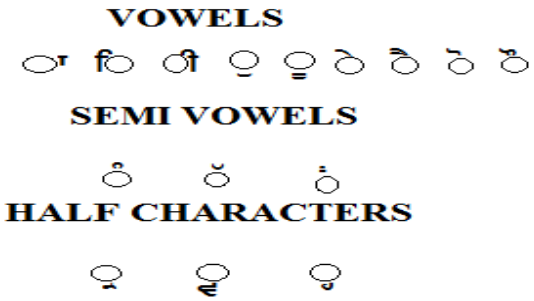

Fig 4: Vowels, Semi Vowels and Half Characters

III.FEATURE EXTRACTION METHODS

It is one of the fundamental key to correctly recognize a text. It can help in better recognition. In this procedure we select a set of specific features by which we can uniquely distinguish each character segmented. There are different feature extraction methods yet determination of a suitable strategy is one of the critical tasks which influence the recognition rate of the framework. Probably the most well-known feature extraction methods are based on the structural and statistical features of the character extracted. Basically features can be classified widely into two categories:-

A. *Structural Features:* - They usually describe geometrical and topological characteristics of a character. They are Less sensitive to character size and font .It includes various :-

1) *Holes:* - These are the loops present in a character. For e.g. character ਕ in Gurmukhi consists of one single hole.

Fig 5: Character with 1 hole.

2) *Junctions:* - A junction is a point where two or more lines converge. Thus in a character image a junction is a point where there are two or more dark pixels in a neighborhood of a single pixel. In Gurmukhi the character ਹ comprises of one junction i.e. with the headline.

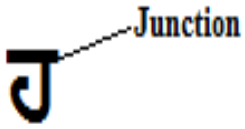

Fig 6:- Character with 1 junction

3) *End points:* - The end points can be defined as the beginning and ending of the line. For e.g. the character ਟ in Gurmukhi consists of 3 end points.

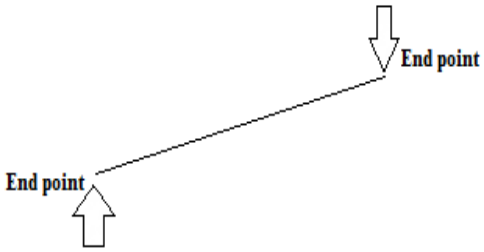

Fig 7: - Endpoints of a line

B. *Statistical Features:* - These statistical features can be acquired by performing some computational operations on character image. This feature set incorporates features like zoning, projection, profiling, histogram and distance, moments etc.

1) *Zoning:* - In this the character image is divided into small windows of mxn size to create individual zones. Native features are perceived from each zone which includes density value, directional features etc. where density is the ratio of total no. of the foreground pixels to the total no. of pixels in a single specific zone [13]. These density values computed from are utilized as feature vectors for the classification purpose.

2) *Projection Histogram:* - This method counts the no. of pixels that are prevailing in precise direction. Direction may be horizontal, vertical, left diagonal and right diagonal [13]. After that corresponding histograms are plotted.

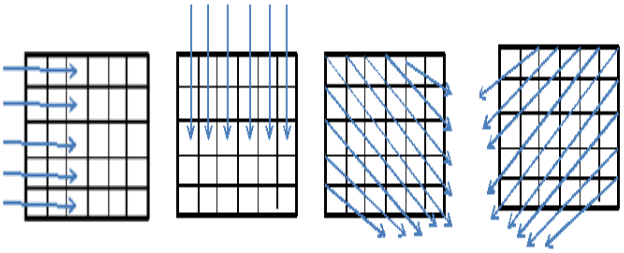

Figure 8: -a) Horizontal projection histogram, b) Vertical projection histogram, c) Diagonal-1 histogram, d) Diagonal-2 histogram.

3) *Profiles*: - In this method no. of pixels are counted between the bounding box of the character and the external edge of the character [13]. This method depicts the external shapes very efficiently allows to distinguish between certain characters that are topologically analogous for e.g. ਸ and ਮ, ਧ and ਪ. These profiles can be defined in four directions left and right projection profile, Right profile depth, left profile lower depth, left profile upper depth [12].

IV. CLASSIFICATION METHODS

This phase is also known as the decision making phase of the recognition process. The features we extracted in the preceding stage are used to classify each segmented character according to the pre-defined rules. The decision rules are made such that there is less probability of misrecognition. There are different types of classification methods and even there are multiple classifiers that can be applied such as Support Vector Machine, Artificial Neural Network. Combination of such classifier's can also be applied for the classification. Some generally used classifiers are:-

A. K-Nearest Neighbor

The k-nearest neighbor (k-NN) approach performs classification by considering the labeled training points as anchor points in the n-dimensional space, where n is the size of feature. It then calculates the Euclidean distance between the test point and all the reference points in order to find the K nearest neighbors, and then distance obtained is ranked in ascending order and consider the reference points corresponding to the K with smallest Euclidean distances. A test sample is then attributed the same class label as the label of the majority of its K nearest neighbors. Euclidean distance is the straight line distance between two points in n-dimensional space.

B. Support Vector Machine

SVM is based on the statistical learning theory which is based on the supervised learning .It is mostly used as a linear classifier [13].In case of supervised learning system is trained to perform a given task on the given input and output pairs .It usually takes a set of inputs and classify them in one of two distinct classes .It isolates two classes by using a hyper plane between them. Main disadvantage of its statistical learning theory is the estimation of output for those inputs for which the system is not trained .There are different type of kernel functions of SVM such as Linear kernel, Polynomial kernel, Sigmoid kernel and Gaussian Radial Basis (RBF) kernel. Selection of the

appropriate kernel is one of the confines of the SVM. It maps the points of the different categories from n – dimensional space into higher dimensional space where these categories are more separable. It tries to find the finest hyper plane in high dimensional space that will best to separate two categories of points.

C. Artificial Neural Network

It is network initially developed according to the operation of human neural system .Due to its high speculation capability and insensitivity to noise it provides an encouraging result in the field of pattern recognition [13]. ANN is composed of three layers of units namely:-

- input layer,
- Hidden layer and
- Output layer.

These units work collaboratively to solve specific issues. It has a training mode and the testing mode .In training mode neural network is trained for s particular pattern and in testing mode when the target pattern is received as the given input then the output turns into the current output .Each input signal has weights connected with them [12] .Each of the input information is processed by calculating the weighted sum of the given inputs .Neural Network can be classified as:-

- Feed forward network
- Feed backward network

V. COMPARISON OF RESULTS

Table below provides an overview of numerous techniques proposed by various researchers.

Sr. No	Title	Publish Year	Description	Advantages	Disadvanta--ges	Technique Used	Results
[1.]	Recognition of Gurmukhi text from signboard images captured by mobile camera	2014	They represented a recognition system for Gurmukhi text from sign board images captured through mobile cameras of different resolutions. They firstly segmented the characters and then feature are extracted to uniquely classify the characters.	1. Use of portable cameras instead of desktop scanners.	1. Upper and lower zone characters are not recognized. 2. Sihari is recognized as kanna ,which leads to misrecognition of text.	Feature Extraction :- Zoning, Classification:-SVM.	Accuracy with:- Linear kernel 92.38%, Polynomial kernel-85.38%
[2.]	Modified Gabor Feature Extraction Method for Word Level Script Identification-Experimentation with Gurumukhi and English Scripts.	2013	They represented a system for the identification of English and Gurmukhi script at word level. This system also identifies English numerals.	1. A multilingual system which recognizes Gurmukhi text from English script and numerals. 2. Improvement over the Gabor feature extraction method without zoning.		Feature extraction:- zone based Gabor filters. Classification:-SVM	Linear kernel:-92.87%, polynomial kernel:-93.28%,; RBF kernel:-99.39%
[3].	Script Identification of Pre-Segmented Multi-Font Characters and Digits	2013	This paper represents a recognition system for pre segmented multifont and multisized Gurmukhi and English characters and numerals .It also provides a comparative study of Gabor and Gradient features using different kernels of SVM.	1. High recognition accuracy of 99.19% is achieved for fonts that were not present in the training set.		Feature extraction methods:- Gabor features Gradient feature. Classification Technique:-Multi class SVM	Average recognition rate with:- 1.Gabor features:-98.9% 2.Gradient features:-99.45%
[4].	Extract the Punjabi Word with Edge Detector from Machine Printed Document Images.	2013	In this they performed two functions of line and letter crop for segmentation and then the extracted characters are compared with the existing database. If the character matched then it is displayed otherwise maximum matched character is displayed. It represents an edge base method.			Sobel edge detector.	
[6]	Identification of Printed Punjabi Words and English Numerals Using Gabor Features.	2011	This paper facilitates multilingual OCR, It represents a system to recognize English numerals and Punjabi words from scanned documents.	1. Facilitates multilingual character recognition	1.This system is restricted to AnmolLipi and Anmol Kalmi font for Punjabi words and Times New Roman and Calibri for English Numerals.	Feature Extraction method:-Gabor Filter. Classification Method:-SVM	Average accuracy with:- Linear kernel:-99.75%, Polynomial kernel:-99.86%, RBF kernel:-96.68%

[7]	Optical Character Recognition of Gurmukhi Script Using Multiple Classifiers	2009	The proposed recognition system is based on 4 classifiers that operate in the serial and parallel mode. For combining the results of these classifiers corpus based weighted voting method is used.	1. Use of multiple classifiers results in better performance as compared to the use of single classifier. 2. Problem of broken characters has also been solved.		1.Binary Tree classifier using:- Structural features 2. The KNN classifier operates on the structural features 3.SVM using Gabor features. 4.SVM using:- Statistical features.	Recognition accuracy by combining the results of all four classifiers is :- 99.59%.
[8]	Degraded Text Recognition of Gurmukhi Script.	2008	This paper provided system for the recognition of degraded text documents. There are several kind of degradations found in the documents like broken characters ,touching characters ,heavily printed characters .The proposed system provides a new approach for degraded document recognition of Gurmukhi script containing heavily printed characters and touching characters.	1. .Recognition of degraded documents.	1. Recognition of touching and heavily printed documents only.	Feature extraction methods:- 1.Structural features. 2.Statistical features: Classification Technique:-K-NN,SVM, Artificial Neural Network.	Recognition accuracy with:- 1.K-NN:-86.09% 2.SVM:-92.54% 3.Neural network:-76.62%
[9].	Structural Features for Recognizing Degraded Printed Gurmukhi Script	2008	This paper represents a work on the recognition of the degraded documents.	Recognition of degraded text.		Feature extraction method:-Structural features. Classification technique:-SVM, K-NN.	Average accuracy with:- 1.SVM:-91.54% 2.K-NN:-83.60% (k=1).
[10].	A Complete Machine printed Gurmukhi OCR System	2006	This paper provides a solution to the various problems encountered during the recognition of Gurmukhi script. Post-processing is also carried out with Punjabi corpus.	1 .Multifont Gurmukhi character recognition system. 2. Eliminates illegal character combination. 3. Provides spell checking.	1. Performance degradation on low quality on low quality text.	Feature extraction methods:- 1.Structural 2.Statistical Classification Technique:-1.Binary tree. 2.Nearest Neighbor	Accuracy above 97%
[11.]	A Gurmukhi Script Recognition System	2000	This represents a recognition system for machine printed Gurmukhi script. Four major fonts Punjabi, Gurmukhi ,Amrit-Lipi, PN-TTAmar and font size of 12,16,20 and 26 is considered.	1 .Need to perform testing on certain sub set of classes thus computation speed increases.	1. Limited to certain fonts and size.	Feature extraction :- 1.Structural features 2. Statistical Features. Classification :- 1.Binary Tree Classifier 2. Nearest Neighbor Classifier.	Recognition rate of 96.6%
[12]	Feature Extraction and Classification for OCR of Gurmukhi Script.	1999	In this they first grouped the characters into 3 subsets depending on their zonal position. Middle zone characters are distributed to subsets using binary decision tree.	1. Testing to be performed only on certain subset classes. 2.Improved computational speed. 3.Hybrid classifier combines strength of binary tree and Nearest Neighbor classifier.	Performed on limited fonts and size.	Feature extraction methods:- 1.Structural features. 2.statistical features. Classification Technique: Nearest Neighbor, Binary tree classifier.	

VI CONCLUSION

This paper represents various OCR systems for machine printed Gurmukhi character and word recognition. We concluded that different recognition models have been put forth in recent years and different research groups are working on the recognition of Gurmukhi characters and words but most of the reported work is on scanned documents only less amount work is accessible on recognition of characters and words on images captured by mobile camera. Reported work on images has an accuracy of 92.38% with linear kernel and 85.38% with polynomial kernel and have some limitations. In future these limitations can be resolved to some degree by using different classifiers and feature extraction methods.

REFERENCES

[1] Shilpa Arora ,Dharamveer Sharma and SilkyArora," Recognition of Gurmukhi text from signboard images captured by mobile camera "International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 17 (2014), pp. 1839-1845.

[2] Rajneesh Rani, Renu Dhir and Gurpreet Singh Lehal," Modified Gabor Feature Extraction Method for Word Level Script Identification- Experimentation with Gurumukhi and English Scripts", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.6, No.5 (2013), pp.25-38.

[3] Rajneesh Rani,,Renu Dhir and Gurpreet Singh Lehal" Script Identification of Pre-Segmented Multi-Font Characters and Digits", 12th International Conference on Document Analysis and Recognition 2013.

[4] Gaurav Singla,Dr.Parmod Kumar," Extract the Punjabi Word with Edge Detector from Machine Printed Document Images", International Journal of Computer Science & Engineering Technology (IJCSET) Vol. 4 ,No. 05 May 2013 ISSN : 2229-3345 .

[5] Usha Rani, Er. Balwinder Singh and Er. Ravinder Singh," Machine Printed Punjabi Character Recognition Using Morphological Operators on Binary Images", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 3, May - 2012 ISSN: 2278-0181.

[6] Rajneesh Rani, Renu Dhir, and G.S. Lehal," Identification of Printed Punjabi Words and English Numerals Using Gabor Features.", World Academy of Science, Engineering and TechnologyVol:5 2011-01-21.

[7] Gurpreet Singh Lehal,". Optical Character Recognition of Gurmukhi Script Using Multiple Classifiers", .In Proceedings of the International Workshop on Multilingual OCR, MOCR '09, 2009.

[8] Manish Kumar, "Degraded Text Recognition of Gurmukhi Script ",Doctor of Philosophy Thesis,Thapar University,2008[online].

[9] M. K. Jindal ,R. K. Sharma and Gurpreet Singh. Lehal, "Structural Features for Recognizing Degraded Printed Gurmukhi Script".

[10] Gurpreet Singh and Chandan Singh,"A Complete Machine Printed Gurmukhi OCR System",Vivek,vol16(3),pp.10-17,2006.

[11] Gurpreet Singh Lehal and Chandan Singh, "A Gurmukhi script recognition system", *Proceedings 15th International Conference on Pattern Recognition*, Barcelona, Spain, Vol. 2, pp. 557-560, 2000.

[12] Gurpreet Singh Lehal and Chandan Singh, "Feature extraction and classification for OCR of Gurmukhi script", *Vivek*, Vol. 12, No. 2, 1999.

[13] Anoop Rekha, "Offline Handwritten Gurmukhi Character and Numeral Recognition using Different Feature Sets and Classifiers - A Survey", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 3, pp. 187-191, May-Jun 2012.