Classification of Printed and Handwritten Gurmukhi text using labeling and segmentation technique

Jaswinder Kaur

Guru Kashi University,TalwandiSabo,Punjab,INDIA

Abstract- This document is based on character recognition which include handwritten and machine written texts. The basic aim of this research is to differentiate between printed or machine text and hand written text. In current work the optical character reader (OCR) for Punjabi Gurumukhi text is performed. Process of research is follow image acquisition, preprocessing, segmentation, feature extraction, classification &recognition, post processing.

Index Terms- OCR, Machine written text, Handwritten text

I. Introduction

In the era of computer, it has become mandatory to have all the available information in a digital form which can easily recognized by machines. However, on the floor of computer data room, there are main two types of data, like as printed material & handwritten script In this research, the deep focus of current research is to draw discriminate printed and handwritten Gurmukhi text is part of hard & soft copies. Character Recognition (OCR) is the process of converting scanned images of machine printed or handwritten text into a computer processable format.

Classification of character recognition system is based on character type and recognition mode. Character Type is further divided as Printed/type-written and Hand-written Character Recognition. Handwritten text & printed text.Handwritten can be further divided into two categories: Cursive and Hand printed script. Recognition of handwritten characters is a much more difficult problem. Characters are non-uniform and can vary greatly in size and style.However printed text includes the materials such as books, newspapers, magazines, documents and various writing units in the video or still image. Machine printed characters are uniform in height,width and pitch assuming the same font and size are used. Problem related to these are solved with little constraint.

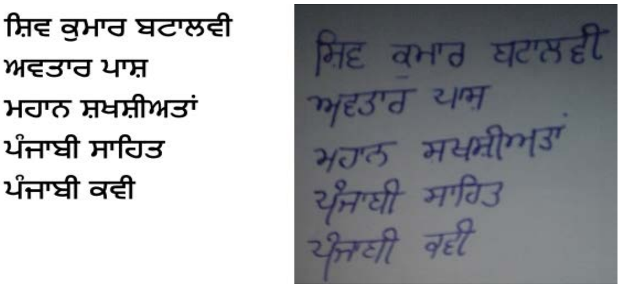

ਸ਼ਿਵ ਕੁਮਾਰ ਬਟਾਲਵੀ
ਅਵਤਾਰ ਪਾਸ਼
ਮਹਾਨ ਸ਼ਖਸ਼ੀਅਤਾਂ
ਪੰਜਾਬੀ ਸਾਹਿਤ
ਪੰਜਾਬੀ ਕਵੀ
ਸਿਵ ਕੁਮਾਰ ਬਟਾਲਵੀ
ਅਵਤਾਰ ਪਾਸ
ਮਹਾਨ ਸਖਸੀਅਤਾਂ
ਪੰਜਾਬੀ ਸਾਹਿਤ
ਪੰਜਾਬੀ ਕਵੀ

Fig.1 Representation of printed and handwritten Gurmukhi text.

There are number of difference between hand written and machine written text.Theseare given below:

- In machine, written language has less no. of strokes as compared to hand written text.
- Machine written text is more readable than the hand-written text.
- Machine written language has fixed sized text compare to the hand-written text.
- Machine written text has more options for formats compare to the hand-written text.
- Machine written text is equally spaced compare to the hand-written text.

II. STRUCTURE OF GURMUKHI CHARACTER

Writing style of Gurmukhi script is from top to bottom and left to right.In Gurmukhi script , there is no case sensitivity. Regular Gurmukhi words can typically be divided into three strips: top, core or middle and bottom as shown in Figure. The header line separates the top strip and the core strip and base line separates core strip and lower strip. The top strip generally contains the top modifiers, and bottom strip contains lower modifiers.

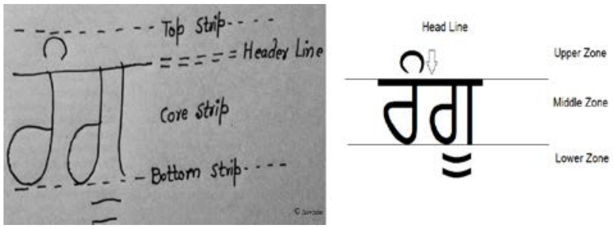


Fig.2 Representation of three Strips of a word in Gurmukhi script.

Gurmukhī has thirty-eight consonants (*akhar*), 10 vowel symbols (*lāgamātrā*), two symbols for nasal sounds (*pair bindi* and *ṭippī*), and one symbol which duplicates the sound of any consonant (*addak*). In addition, four conjuncts are used: three subjoined forms of the consonants Rara, Haha and Vava, and one half-form of Yayya. Use of the conjunct forms of Vava and Yayya is increasingly scarce in modern contexts.

I. Proposed System :

There are number of techniques for recognition of characters.Mostly methods are too much close to each other.In this thesis I follow these steps to find out the result-imageacquisition,preprocessing,segmentation, feature extraction, classification & recognition, post processing.

A.Image acquisition

The images are acquired through the scanner. The images are of RGB in nature (Colored).If we deeply observed any image then we can find that there are a number of unwanted elements that can be considered as noise. These elements can create difficulties in the way of performance of the whole system. So that it becomes necessary to remove this noise. Preprocessing works on it.

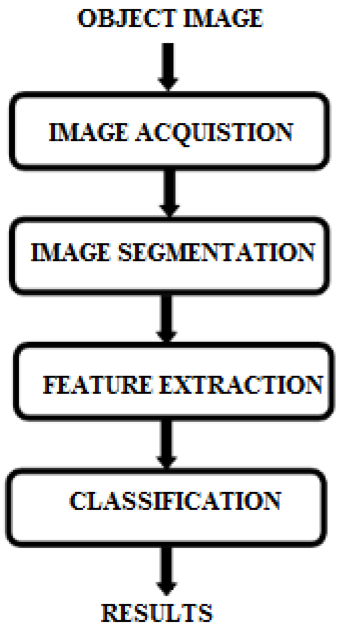


Fig.3 Representation of Proposed System

B. Image segmentation

Segmentation means partitioning of image into various part of same features or having some similarity. In the other words,Segmentation is the phase in which data is decomposed at character or stroke level so that nature of each character or stroke can be studied individually. The segmentation can be done using various methods like otsu' method, k-means clustering, converting RGB image into HIS model, cross correlation etc.
Segmentation is carried out mainly two stages namely – Line segmentation and Word segmentation.

1) *Line Segmentation:*The line segmentation is carried out by scanning the entire row one after the other and provide results.

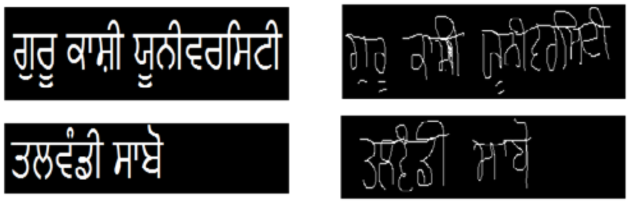


Fig.4 (a) Line segmented printed text (b) Line segmented handwritten text

2) *Word segmentation:* In it, the same principle used in line segmentation is used. The only difference here is that the scanning process is carried out vertically.

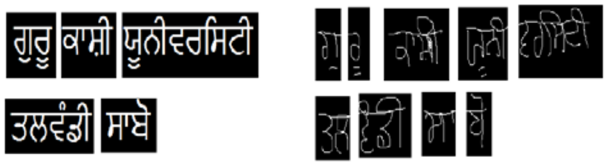


Fig.5.a)Word segmented printed text (b) Word segmented handwritten text

Otsu Threshold Algorithm for segmentation: Thresholding creates binary images from grey-level images by setting all pixels below some threshold to zero and all pixels above that threshold to one. The Otsu algorithm defined in is as follows:
i) According to the threshold, separate pixels into two clusters
ii) Then find the mean of each cluster.
iii) Square the difference between the means.
iv) Multiply the number of pixels in one cluster times the number in the other

a) *C. Feature Extraction*

b) Feature extraction is used to extract relevant features for recognition of characters based on these features. First features are computed and extracted and then most relevant features are selected to construct feature vector which is used eventually for recognition. The computation of features is based on main two types of feature

c) extraction:

- Statistical feature extraction
- Structural feature extraction

a) *Statistical feature extraction :* In this type of extraction the extracted feature vector is the combination of all the features extracted from each character. The associated feature in feature vector of this type of extraction is due to the relative positions of features in character image matrix.

b) *Structural feature extraction :* This is a primitive method of feature extraction which extracts morphological features of a character from image matrix. It takes into account the edges, curvature, regions, etc. This method extracts the features of the way character are written on image matrix.

d) The functions that are used in feature extraction are:

e) *Indexing and labelling :*This is a process by which distinct characters in an image are indexed and labelled in an image. Thus helps in classification of characters in image and makes feature extraction of characters simple.

f) *Boxing and Cropping :* This is a process of creating a boundary around the characters identified in an image. This helps by making cropping of characters easier. After boxing the characters are cropped out for storing them as input variables for recognition.

g) *Reshaping and Resizing :* Reshaping is done to change the dimensions of the acquired character in desired shape. Resizing is done to reduce the size of characters to a particular minimum level

h) *D.QUALITY METRICS*

The results achieved from segmentation depend upon the quality of the images. The images get distorted during the acquisition process which degrades its quality. The quality of the image can be judged by the human eye but the results of the segmentation cannot be judged by visualizing. Also in real time environment the subjective evaluation of segmentation gets complex. Therefore there requires some standard parameters for performance evaluation.

(i) SSIM (Structural Similarity Index Measure): It is well known quality metric that is used to calculate the similarity between two images. It is designed by modeling any image distortion as three factors that are contrast distortion, luminance distortion and correlation.

(ii) MSE (Mean Square Error): It is a method used to check for errors. Two MSEs are calculated and then are compared to find the accuracy of an image. It calculates the quantitative score that helps to measure the degree of homogeneity or the level of error or distortion between them. When a zero-mean random source x passes through a cascade of K additive independent zero means distortions. A lower MSE value will result in higher quality image.

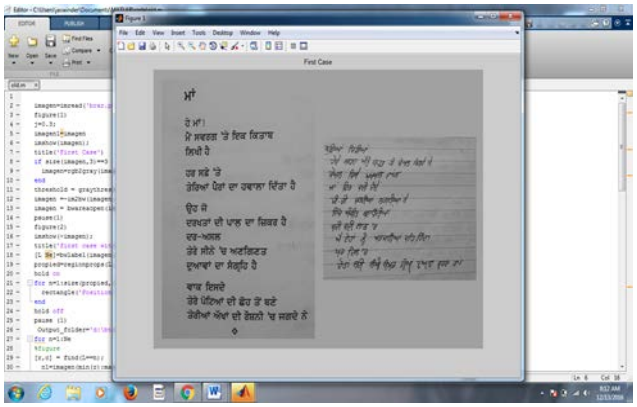
*(iii)PSNR(Peak Signal Noise Ratio):*The PSNR is most commonly used as a measure of quality of reconstruction of loss compression codecs. The signal in this case is the original data, and the noise is the error introduced by compression. A higher PSNR would normally indicate that the reconstruction is of higher quality. Performance Evaluation of Preprocessed Images using PSNR, MSE and Mutual Information metrics.

III. Results & Discussion

In this research work,I have taken the Punjabi (Gurmukhi) alphabets for both hand written and machine written. Our objective is to identify and segment the letters from the image containing text. For it we will take various parameters to have automatic differentiation between machine written and hand written text. These parameters are aspect ratio and wavelength value. Machine written text has constant aspect ratio and common wavelength. But hand written text has varying aspect ratio and the wavelength.

The images with both types of text like hand written and machine written text are given as input. With the help of threshold based technique the text characters are recognized such that automatic system can be developed which can recognize the hand written text and machine written text.With the help of powerful software Matlab,ihave done a number of experiments on the handwritten and machine written documents.Here I have a experiment with result.

EXPERIMENT



```
imagen=imread('brar.p
figure(1)
j=0.3;
imagen1=imagen
imshow(imagen);
title('First Case')
if size(imagen,3)==3
  imagen=rgb2gray(ima
end
threshold = graythres
imagen =~im2bw(imagen
imagen = bwareaopen(i
pause(1)
figure(2)
imshow(~imagen);
title('first case wit
[L Ne]=bwlabel(imagen
propied=regionprops(L
hold on
for n=1:size(propied,
  rectangle('Position
end
hold off
pause (1)
 Output_folder='d:\b
for n=1:Ne
%figure
[r,c] = find(L==n);
  n1=imagen(min(r):ma
```

ਮਾਂ

ਹੇ ਮਾਂ!
ਮੈਂ ਸਫ਼ਰ 'ਤੇ ਇਕ ਕਿਤਾਬ
ਲਿਖੀ ਹੈ

ਹਰ ਸਫ਼ੇ 'ਤੇ
ਤੇਰਿਆਂ ਪੈਰਾਂ ਦਾ ਹਵਾਲਾ ਦਿੱਤਾ ਹੈ

ਉਹ ਜੋ
ਦਰਖ਼ਤਾਂ ਦੀ ਘਾਲ ਦਾ ਜ਼ਿਕਰ ਹੈ
ਦਰ-ਅਸਲ
ਤੇਰੇ ਸੀਨੇ 'ਚ ਅਣਗਿਣਤ
ਦੁਆਵਾਂ ਦਾ ਸੰਗ੍ਰਹਿ ਹੈ

ਬਾਕ ਇਸਦੇ
ਤੇਰੇ ਪੋਟਿਆਂ ਦੀ ਛੋਹ ਤੋਂ ਬਣੇ
ਤੇਰੀਆਂ ਅੱਖਾਂ ਦੀ ਰੋਸ਼ਨੀ 'ਚ ਜਗਦੇ ਨੇ

Fig.6 Input Image

After giving the input, the characters are recognized. Image after character recognition are shown below:

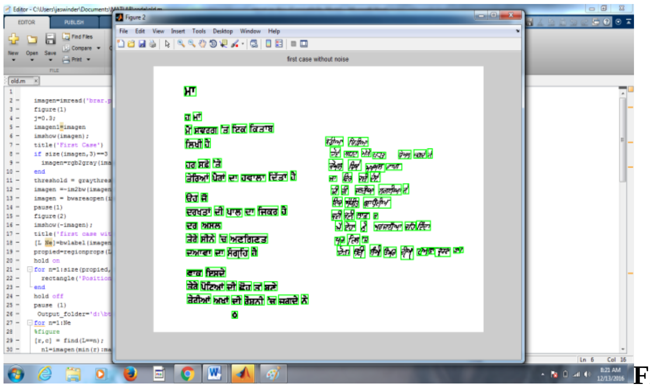


Fig.7

Image after character recognition

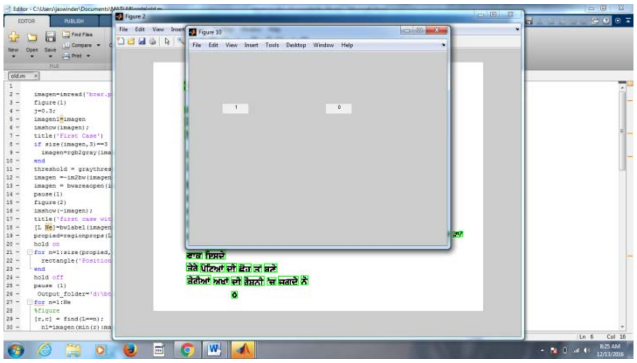


Fig.8 Output Image

Machine written text is shown as 1 and hand written text is shown as 0.

References

[1] Abhishek Jindal, and Mohd Amir. **"Language Independent Rule Based Classification of Printed & Handwritten Text"** in Proceedings of the IEEE 2015 tenth workshop on Multimedia Signal Processing, pp.393-398, 2015.

[2] B. B. Chaudhuri. "Automatic Separation of Machine-Printed and Hand-Written Text Lines " Pattern Recognition Letters ,2015.

[3] ErginaKavallieratou, and StathisStamatatos. "Discrimination of Machine-Printed from Handwritten Text Using Simple Structural Characteristics" in Proceedings of the IEEE 2004.

[4] MehryarEmambakhsh, Yulan He, Ian Nebney."Handwritten and Machine-Printed text Discrimination using a template approach ". Proceedings of the IEEE 2014.

[5] PurnenduBanerjee, and A Nikolaidis. "A System for Hand-Written and Machine-Printed Text Separation in Bangla Document Images" in Proceedings of the IEEE 2012.

[6] RanjeetSrivastava, and Ravi Kumar Tewari. **"Separation of Machine Printed and Handwritten Text for Hindi Documents"**in Proceedings of the IEEE 2015.

[7] Mrs.Saniya Ansari. "Optimized and Efficient Feature Extraction Method for Devanagari Handwritten Character Recognition", Pattern Recognition, Letters.

[8] SurabhiNarayan, and A Nikolaidis. **"Discrimination of handwritten and machine Printed text is Scanner document Images based on Rough Set Theory"** in Proceedings of the IEEE 2012 tenth workshop on Multimedia Signal Processing, pp.393-398, 2012.

[9] TanzilaSaba, and A Nikolaidis. **"Language Independent Rule Based Classification of Printed & Handwritten Text"** in Proceedings of the IEEE 2015 tenth workshop on Multimedia Signal Processing, pp.393-398, 2015.

[10] U.Pal."Machine-printed and hand-written text lines identification" in Proceedings of the IEEE 2001.

AUTHORS

First Author – Jaswinder Kaur, Guru Kashi University,TalwandiSabo,Punjab, INDIA
jasharmalke@yahoo.in