ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 5; Issue 3; May-June: 2017; Page No. 49-54
Journal Approved by UGC

Comparative Analysis for Gurmukhi and Devanagari Script at Word Level

Sapna Dhiman[1], Rohit Sachdeva[2], Sumeet Kumar[3], Neha Khanna4

[1]Assistant Professor, M. M. Modi College, Patiala

dhiman.sapna@gmail.com

[2]Assistant Professor, M. M. Modi College, Patiala

rsachu.147@gmail.com

[3]Assistant Professor, M. M. Modi College, Patiala

ksumeet2012@yahoo.com

[4]Assistant Professor, Govt. Bikram College of Commerce, Patiala

Abstract

This paper presents a comparative analysis for Gurmukhi OCR with Devanagari OCR at word level. As OCR works in different stages and each stage has its own importance. In this paper, feature extraction and classification methods are discussed. Features are the set of minimum values of images to describe it uniquely. Before the classification, Features have been extracted from word images. In this paper, two feature extraction techniques Discrete Cosine Transform (DCT) and Gabor filter has been used. Gabor produces 189 features and DCT produces 100 features in zig-zag method. For training and testing, 50 different classes with 30-35 samples of each class for training and 10-15 samples for testing have been taken in Gurmukhi Script as well as Devanagari Scripts. For classification k-NN classifier with value of k=3, 5 and SVM have been used for performance comparison.

Keyword: Feature extraction, Gabor Filter, DCT, k-NN Classifier, OCR.

1. INTRODUCTION

Optical Character Recognition is a technique used to digitize the machine printed or hand written data. Lots of data and information is available around us in the form of printed pages, newspaper, ancient books. But it is difficult for a person to extract the information according to his need from these sources. So need a source which helps in getting this information. OCR helps to get that information in digital form so that everyone can access that easily. OCR falls under two main categories:

a. Machine Printed Recognition
b. Handwritten Text Recognition

Developing a Handwritten Text Recognition system is more difficult than Machine Printed Recognition system, because of variation in writing styles and methods [1,2]. Many good quality OCRs are available in different Indian Languages like Oriya, Tamil and Telugu etc. Devanagari and Gurmukhi are the most common and popular scripting languages not in India but also in the world [3]. This paper depicts the performance analysis different 50 classes of both languages using common feature extraction techniques and classification method.

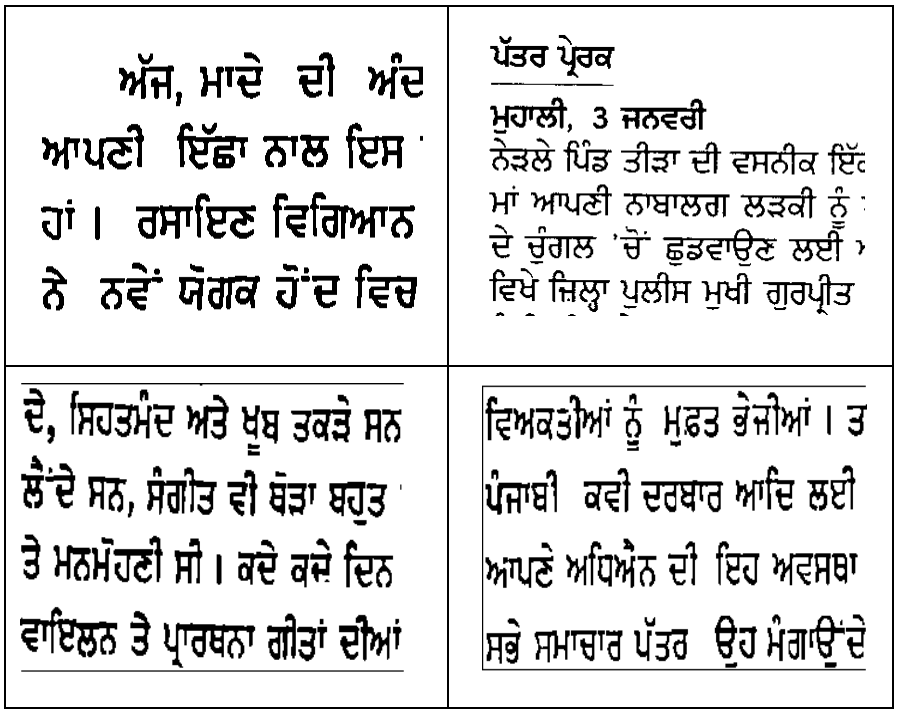
2. Corpus and Binarization

Gurmukhi Script

In case of Gurmukhi OCR, the corpus has been collected from different scanned booksand newspaper at 300 dpi at word level. In Table 1, Samples of scanned books and newspapers are shown:

Corresponding author:Rohit Sachdeva

Table 1: Samples of scanned books

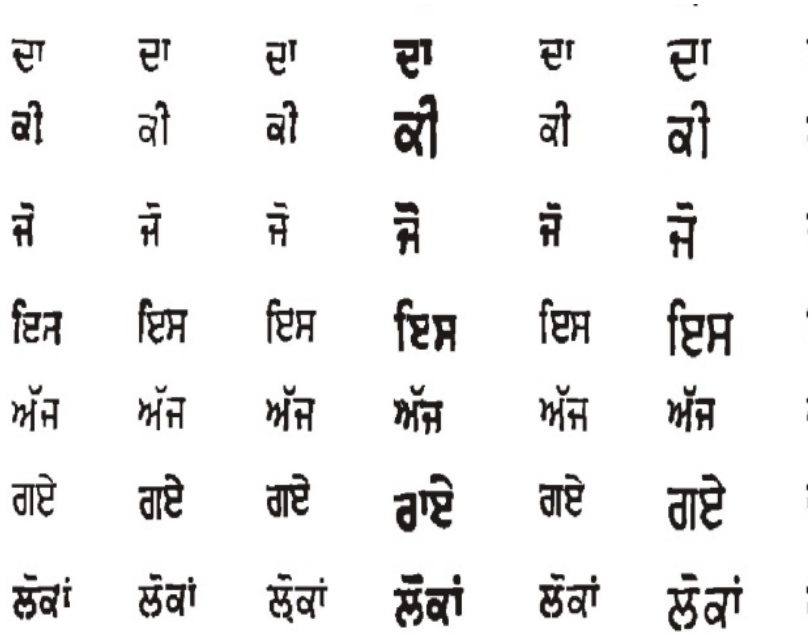


After completing binarization of data next step is to segment the word from scanned images. The segmentation stage has three steps:

- Line segmentation: Where scanned pages are segmented into lines.
- Word segmentation: Where segmented lines are further segmented into word images.
- Character segmentation: where a segmented word is segmented into character level.

Collected word images of Gurmukhi are shown in Table 2:

Table 2: Corpus of segmented word images



Devanagari Script

For Devanagari OCR, corpus is collected as handwritten text in forms. For that purpose, a template is generated and filled by different persons to have a variation in writing styles and samples. Sample of form is shown Figure 1:


Data Collection form for Devanagari ICR and Form Processing Research
Kindly use your natural handwriting for filling the fields of the form in Hindi Only

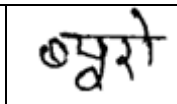
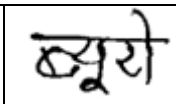
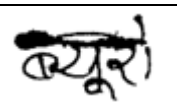
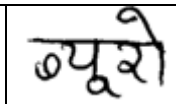
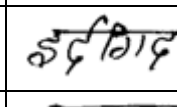
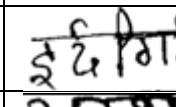
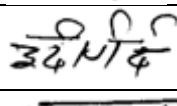
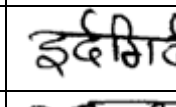
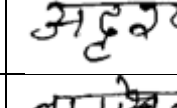
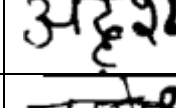
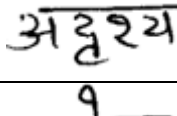
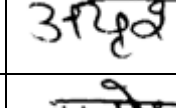
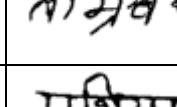
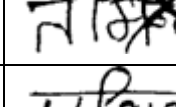
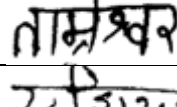
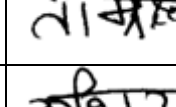
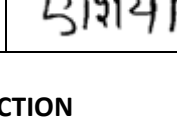
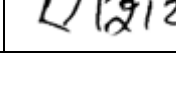
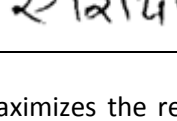
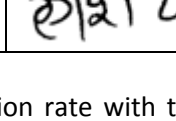
निम्नलिखित शब्दों को डिब्बो में लिखों

द्वेष		ब्यूरो		सभ्य		अकूल		खरबूज़ा		ग्रहण		ताम्रेश्वर	
कृतघ्न		ईंधन		सुफ्फा		अय्यूब		अँगड़ाई		इर्दगिर्द		आज्यपात्र	
द्रव्य		लक्ष्य		छप्पर		घटना		अदृश्य		विलुप्त		प्रेरणास्रोत	
क्षण		फ्रांस		अच्छा		त्रिस्टी		ऊर्मिला		एशिया		अंतर्राष्ट्रीय	
व्रत		शुद्धि		कंट्रोल		विक्ख		गुरुभाई		ऋग्वेद		एक्सप्रेस	
सत्य		स्थल		ट्रैफ़िक		ओद्रनी		विभाज्य		पथभ्रष्ट		अश्विन्दर	
कथ्य		श्रद्धा		द्विज		अक्षर		उत्तेजित		ह्यात		कुरुविल्ला	
क्रोध		इकट्ठा		चंदोक		अरण्य		ज्ञातव्य		मद्देनज़र		झिलमिल	
चोल		कृष्ण		अँग्रेजी		उंगली		फ्रांसिस		उत्तम		कृतघ्य	
मिट्टी		काम्या		ध्यान		चिह्न		वज्रहस्त		ठिल्ला		पारब्रह्म	

Figure 1: Sample of form for data collection

These forms have been filled by different persons. After preprocessing and binarization of forms, text has been extracted from these forms [4]. Some collected samples are shown in Table 3:

Table 3: Samples of handwritten text of Devanagari

ब्यूरो				
इर्दगिर्द				
अदृश्य				
ताम्रेश्वर				
एशिया				

3. FEATURE EXTRACTION

In any OCR system, feature extraction is most important and necessary step. The result of classification stage depends upon the extracted features. The major goal of the feature extraction stage is to find and extract such features of images, which maximizes the recognition rate with the least amount of values. So before choosing any classification method, it is important to decide which feature technique is best for the system and which is not. In this paper, Gabor and DCT features have been used for feature extraction of images.

A) GABOR FILTER

It is a very common and widely used method applied on images [5]. A Gabor filter is very popular in face recognition, texture and character recognition [8, 9]. A Gabor filter is selective to both spatial frequencies as well as orientation frequency so sometimes called as a kind of local narrow band pass filter. The equation of 2D Gabor filter is given below:

$$f(x, y, \phi, \sigma_x, \sigma_y) = \exp\left[-\frac{1}{2}\left\{\frac{R_1^2}{\sigma_x^2} + \frac{R_2^2}{\sigma_y^2}\right\}\right] \times e\left\{i\frac{2\pi R_1}{\lambda}\right\}$$

where $R_1 = xcos\phi + ysin\phi$ and $R_2 = -xsin\phi + ycos\phi$

σ_x and σ_y are the standard deviations of Gaussian envelop along x-axis and y-axis but here $\sigma_x = \sigma_y$ And λ and ϕ are the wavelength and orientation of plane waves.

φ is an angle to rotate the x-y plane to get different orientation values. The value of φ is given by

$\phi = \pi(k-1)/m$,, here k = 1, 2, ….m

where m denotes the number of orientations. Here m=9, so 189 features are extracted from images.

B) DCT

DCT is a most widely used and powerful transform for extracting the features. It is the member of a family of sinusoidal unitary transforms [3, 5], which encodes the significant details or energy or frequency of the image in a few coefficients very efficiently [7]. These transformed coefficients are used as features of the sample image. It calculates the two-dimensional cosine transform an image. The equation of DCT image has been represented as:

$$D(i,j) = c(i)c(j)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} p(x,y)\cos\left[\frac{(2x+1)}{2M}i\pi\right]\cos\left[\frac{(2y+1)}{2N}j\pi\right]$$

Where:

$$C(i) = \begin{cases}\sqrt{\frac{1}{M}}, & if\ i = 0\\ \sqrt{\frac{2}{M}}, & if\ i > 0\end{cases} \text{ and } C(j) = \begin{cases}\sqrt{\frac{1}{N}}, & \text{if } j = 0\\ \sqrt{\frac{2}{N}}, & \text{if } j > 0\end{cases}$$

Here M and N are the height and width of the image. We have scaled the images into 40*40 size, so in this function M=N.

D(i, j) represents the DCT coefficient of the image corresponding to pixel p(x, y).
These coefficients are named as DC component, which is the first coefficient i.e at [0, 0] and AC component, which are the rest of the coefficients of the image. Total 1600 features can be obtained. But we have picked only 100 features, which are selected in a zigzag manner as shown in Fig. 2:

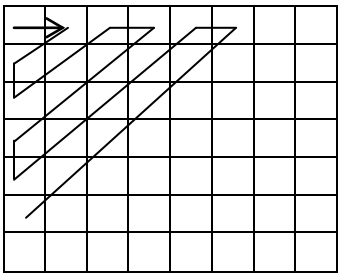


Figure 2: Coefficients selection in zigzag

Table 4 depicts the features and their vector size, which has been used further in classification stage. These feature vectors play important role in classification, as they are used as input to train the OCR and test in the recognition process. Both DCT and Gabor Feature vectors have given in table 4:

Table 4: Features and their vector sizes

S. No.	Feature	Size
1	Gabor Filter	189
2	DCT	100

4. CLASSIFICATION

The classification is important and decision making stage of any recognition system. It assigns the input features of stored pattern and compares it to find out best matches. We have used k-NN classifier for recognition of word images.

k-NN Classifier

In the testing stage, k- nearest neighbor (k-NN) classifier is popular and simplest classifier [6]. First, the system is trained with some samples. It simply

stores training samples with its label. For prediction of a sample, its distance is computed from training sample. After computing distance, the k closest training samples are kept, where k is a fixed integer having value . After that a label is searched. This searched label is a most common label among all those samples, which is the prediction for test sample.

The value of k and the distance function:two major design choices are taken to apply k-NN. In this paper, k = 3 and 5 is chosen for minimum distance. d(x, y) is the distance evaluated between training and test sample, which is computed as:

$$d(x,y) = \| x - y \| = \sqrt{(x-y)*(x-y)} = \left(\sum_{i=1}^{m}(x_i - y_i)^2\right)^{\frac{1}{2}}$$

Where $x, y \in \mathbb{R}^m$.

SVM Classifier

Support Vector Machines (SVM) is also known as kernel Method [9]. It is a guided learning machine under the category of machine learning methods. This method can be used for binary classification, or regressions. SVM are based on the Structural Risk Management Principle which tries to curtail an upper bound of the generalization error instead of to curtail the training error. SVM is very effective in high dimensional spaces. It uses a subset of training points in the decision function, which is called support vector, which makes it memory efficient. In this, distinct kernel methods can be determined for the decision function. The different kernel methods used are given in Table 5

Table 5: Different kernel Methods of SVM

Method Name	Formula
Linear kernel	$K(x_i, x_j) = x_i^T x_j$
Polynomial Kernel	$K(x_i, x_j) = (x_i^T x_j + 1)^d$

I. PERFORMANCE COMPARISON

By using DCT and Gabor filter, features have been extracted for all training and testing samples. Using specified k-NN classifier, the system is trained to recognize the word. Table 6 and Table 7 shows the comparative results of recognition for 50 classes in both Gurmukhi and Devanagari Scripts by using KNN and SVM.

Table 6: Performance using KNN with Gabor Filter and DCT

S. No.	Feature	Gurmukhi Text Recognition %		Devanagari Text Recognition %	
		k=3	k=5	k=3	k=5
1.	Gabor filter	92.14	92.62	88.17	89.35
2.	DCT	96.22	96.99	91.36	91.49

Table 7: Performance using SVM with Gabor Filter and DCT

S. No.	Feature	Gurmukhi Text Recognition %		Devanagari Text Recognition %	
		SVM		SVM	
		Linear	Polynomial	Linear	Polynomial
1.	Gabor filter	94.32	94.85	90.89	91.28
2.	DCT	95.71	95.93	89.48	89.65

5. CONCLUSION AND FUTURE WORK

It is clear from table 6 and 7 , that DCT has provided better results for k=5 in the k-KNN classifier and Gabor filter has provided better results with SVM (Polynomial) . Moreover, accuracy in recognition of machine printed text is more than the accuracy of recognizing handwritten text. None of these gives 100% accuracy. Moreover, there will be a scope to

apply more feature extraction techniques to get better performance and also combination of these feature extraction methods and classifiers can be used to increase the performance.

REFERENCES

1. Y. Tawde and M. Kundargi, "An Overview of Features Extraction Techniques in OCR for Indian Scripts Focused of Offline Handwriting", International Journal of Engineering Research and Application, Vol. 3, Issue 1, pp 919-926, 2013.
2. Kunkari, "Optical Character Recognition System for Devanagari Script", International Journal of Innovative Research in Computer and Communication Engineering", Vol. 4, Issue 7, pp 14028- 14033, 2016.
3. Saidas, Rohithram, Sanoj and Manju,"Malayalam Charater Recognition using Discrete Cosine Transform", International Journal of Engineering and Computer Science, Vol. 5, Issue 2, pp 15741-15743, 2016.
4. Rohit Sachdeva and Dharamveer Sharma "Data Extraction from Hand-filled Form using Form Template" International Journal of Recent and Innovation Trends in Computing and Communication, Vol. 3, Issue 8 , pp. 5311-5317, August 2015.
5. Lehal and Singh, "Feature Extraction and Classification for OCR of Gurmukhi Scripts", Vivek, Vol. 12 Issue 2, pp 2-12, 1999.
6. Rajesh Babu, "OCR for Printed Telagu Documents", project report of M.Tech, pp 1-32, 2014.
7. Charan K., "A Block DCT based Printed Character Recognition", a dissertation submitted for Master of Science, pp 1-69.
8. Singh and Lehal, "Comparative Performance Analysis of Feature(S)- Classifier Combination for Devanagari Optical Character Recognition", International Journal of Advanced Computer Science and Application, Vol. 5, Issue 6, pp 7 – 42, 2014.
9. Arya, Chhabra and Lehal, "Recognition of Devanagari Numerals using Gabor Filter", Indian Journal of Science and Technology, Vol. 8, Issue 27, pp 1 – 6, 2015.