ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Performance Comparison of Gurmukhi Script: k-NN Classifier with DCT and Gabor Filter

SapnaDhiman
Department of Computer Science and Management
M. M. Modi College
Patiala, Punjab, India

Dr. Gurpreet Singh Lehal
Department of Computer Science
Punjabi University
Patiala, Punjab, India

Abstract:This paper presents a comparative performance analysis for Gurmukhi OCR at word level. To evaluate the performance k-NN classifier has been used. Before the classification, Features have been extracted from word images. For feature extraction, word images have been scanned and these images are machine printed images.Here Discrete Cosine Transform (DCT) and Gabor filter has been used to extract the features. DCT provides 100 features of scanned images in zig-zag method andGobor provides 189 features for scanned images. To train the classifier of Gurmukhi OCR, 50 different classes with 30-35 samples of each class i.e 1600 samples have been taken. 750 samples have been used to test the system. Using Gabor filter, k-NN classifier provides 92.6229%of correctness while with DCT with k-NN provides 96.9945% of accuracy.

Keyword:Feature extraction, Gabor Filter, Discrete Cosine Transform (DCT), Classifier, k-NN, OCR.

I. INTRODUCTION

Optical Character Recognition is a technique used to digitize the machine printed or hand written data. We have lots of data and information available in ancient books, bibliography or in printed pages. But it is difficult to read and get all information from that. OCR helps to get that information in digital form so that everyone can access that easily. According to working or recognition way, OCR falls under two main categories:

a. Machine Printed Recognition
b. Handwritten Text Recognition

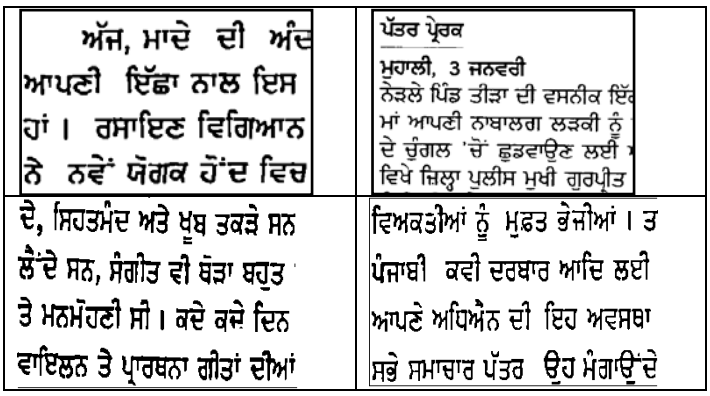
In case of Machine printed recognition system, printed text is scanned with the good quality scanners. The quality of text depends upon the quality of printed text as well as resolution of scanning. Machine printed recognition is easier than handwritten text recognition because there is a lot of variation in writing style[1, 2]. In case of handwritten text recognition, on-line or off-line techniques are used, which make it more complicated.A special pen is used for On-line recognition system and it is a dynamic approach for recognition.A digitizer or a special surface, which senses all movements of pen-tip and also all the movements like pen-up/ pen-down etc. Opposite to that, Off-line recognition system used scanned data, written on paper or any readable material.

II. DATABASE COLLECTION

Lots of work has been done in the field of OCR in many Indian languagese.g Oriya, Malayalam, Telugu, Devanagari etc. Gurmukhi is a popular language not only in India but also in the world as it is 14th most widely spoken language in the world[3]. It is the 14th most widely spoken languages in the world. Gurmukhi has 3 are vowel carriers, 38 consonants, 9 vowels, 3 half vowels, and 3 half characters.A lot of work has been done in Gurmukhi OCR at the character level. Lehal and Singh, describes the features (primary or secondary) of Gurmukhi script[4]. Jindal, lehal and Sinha, describe problem occurred in segmentation and solution for that degraded Gurmukhi text[5]. Lehal,discussed the classification on Gurmukhi characters. How multiple classifiers are used to get better accuracy is also discussed in this paper.But all work is done on at character level[6]. This paper presents the recognition rate of Gurmukhi script at word level.

To enhance existing Gurmukhi OCR, the corpus has been generated from different sources at word level. A scanner is used to transfer the printed text into computer system in digital form. In this paper, images are collected from different machine printed books. These books have been scanned at 300 dpi. Samples of scanned books are shown in table 1:

Table 1: Samples of scanned books



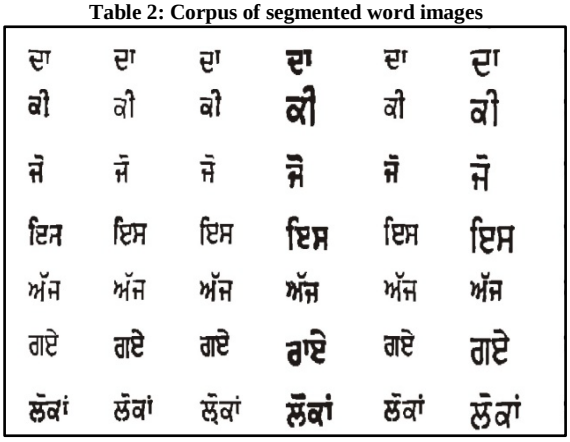
When digitization is completed, preprocessing steps have been applied on these scanned images[3]. Noise removal and binarization have been applied on images.

After preprocessing, next is to segment the pages into word images[7]. 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.

But in this paper, word level images are taken in database collection. Some samples are shown in table 2

Table 2: Corpus of segmented word images



III. EXTRACT FEATURES

The most important step in any OCR system is to extract the features of images. The recognition accuracy of Classifier depends on 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. This stage analyzes the input image and selects a set of features that uniquely classifies the word. To extract the features from scanned images, DCT and Gabor filter have been used.

A) GABOR FILTER

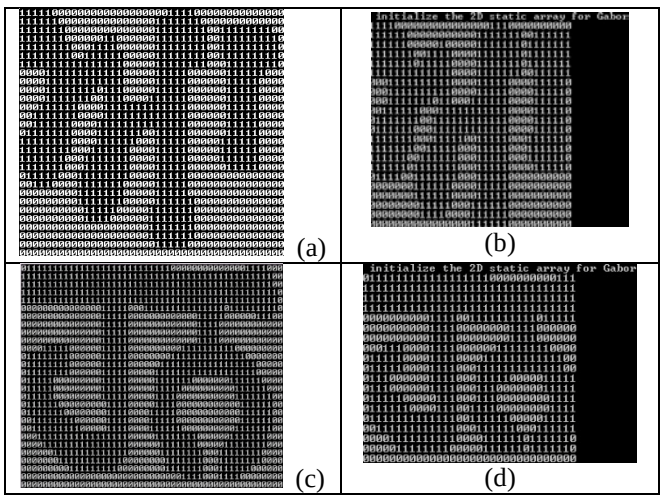
Gabor filter is widely used feature extraction method applied on thescanned images. [10]. A Gabor filter is selective to both spatial frequency as well as orientation frequency so sometimes called as a kind of local narrow band pass filter. A Gabor filter is very popular in face recognition, texture and character recognition [9]. 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 wave.Before extract the features from scanned images, Gabor filter scaled the images into 32*32 matrix.The output 32*32 scaled images for Gabor filter are shown below in table 3,

Table 3: (a), (c) original images and (b), (d) scaled 32*32 array



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

$$\phi = \pi(k-1)/m, \text{where k = 1, 2, …..m.}$$

where m denotes the number of orientations.

By taking m = 9, including all orientations of the whole image as well as taking each quadrant and each sub-quadrant, total 189 features have been extracted from a scanned image.

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], which encodes the significant details or energy or frequency of the image in a few coefficients very efficiently [8]. These transformed coefficients are used as features of the sample image. It calculates the two-dimensional cosine transform an image [9]. The equation of 2D-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:

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

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).

Therefore the coefficient corresponding to all the image pixels will constitute a feature vector set. 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. As the image is scaled to size 40*40, so total 1600 features can be obtained. But we have picked only 100 features, which are selected in a zigzag manner as shown in Fig. 1:

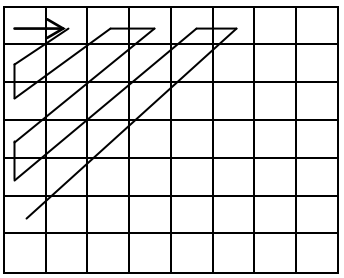


Fig.1: Coefficients selection in zigzag

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 size

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

IV. CLASSIFICATION

The classification is important and decision making stage of any recognition system. It assigns the input features of stored pattern and compare 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. 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 $k \geq 1$. 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 $x, y \in \mathbb{R}^m$.

V. 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 4 depicts the result of recognition for 50 classes.

Table 5: Performance comparison

S. No.	Feature	Recognition %	
		k=3	k=5
1.	Gabor filter	92.144	92.623
2.	DCT	96.223	96.995

As the table 5 indicates DCT with k-NN performs betteras compared to Gabor with k-NN where k=5. Chart 1 depicts the performance analysis of both features.

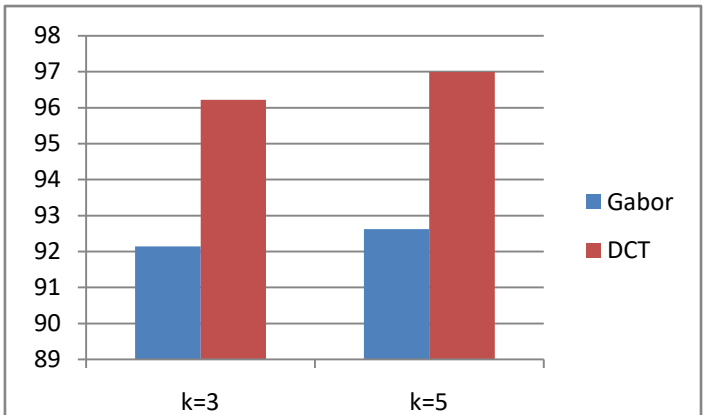


Chart 1: Comparing DCT and Gabor

VI. CONCLUSION AND FUTURE WORK

It is clear from table 4 that DCT has provided better results for k=5 in k-NN classifier with an accuracy of 96.995%. As the analysis has been done with k-NN classifier only, there will be a possibility of testing and training the system with some other classifiers in the future.

Moreover, there will be a scope to apply more feature extraction techniques to get better performance.

VII. 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,"MalayalamCharater Recognition using Discrete Cosine Transform", International Journal of Engineering and Computer Science, Vol 5, Issue 2, pp 15741-15743, 2016.

[4] Lehal and Singh, "Feature Extraction and Classification for OCR of Gurmukhi Scripts", Vivek, Vol. 12, No. 2, pp 2-12, 1999.

[5] Jindal, Lehal and Sinha, "Segmentation Problems and Solutionsin Printed Degraded Gurmukhi Text", International Journal of Signal Processing, Vol 2, Issue 4, pp 258-267, 2005.

[6] Lehal, "Optical Character Recognition of Gurmukhi Script using Multiple Classifiers", proceedings of International Workshop of Multilingual OCR, Article no. 7, Barcelona, Spain, 2009.

[7] Rajesh Babu, "OCR for Printed Telagu Documents", project report of M.Tech, pp 1-32, 2014.

[8] Charan K., "A Block DCT based Printed Character Recognition", a dissertation submitted for Master of Science, pp 1-69.

[9] 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, No 6, pp 7 – 42, 2014.

[10] Arya, Chhabra and Lehal, "Recognition of Devanagari Numerals using Gabor Filter", Indian Journal of Science and Technology, Vol 8 (27), pp 1 – 6, 2015.