INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

FEATURE EXTRACTION FOR GURMUKHI: GABOR FILTER AND DCT

Sapna Dhiman[1]

[1]Assistant Professor of Computer Science Department, M. M. Modi College, Patiala,

dhiman.sapna@gmail.com

Abstract: - In this paper, feature extraction method is described for Gurmukhi optical character recognition system. For feature extraction, word images have been scanned and these images are machine printed images. After preprocessing stage, features are extracted from the scanned images. Here Discrete Cosine Transform (DCT) and Gabor filter are used to extract the features. DCT provides 100 features of scanned images in zig-zag method and Gobor provides 189 features for scanned images. These features further help in classification stages to recognize the word. Feature extraction is every important stage in OCR. The result of classification stage totally depends on the features of images.

Keyword: Feature extraction, OCR, Discrete Cosine Transform (DCT), Gabor filter.

1. Introduction

Optical Character Recognition is widely used technique to digitize the machine printed or hand written data. Lots of work has been done in the field of OCR for many languages [1]. Moreover, OCR is also categorized into two categories:

- Handwritten text recognition
- Machine printed text recognition

Handwritten text recognition is more complicated than text Machine printed text recognition because of variation in writing style, writing ways, etc. Moreover handwritten text recognition system is also categorized in two ways:

- On-line handwritten text recognition
- Off-line handwritten text recognition

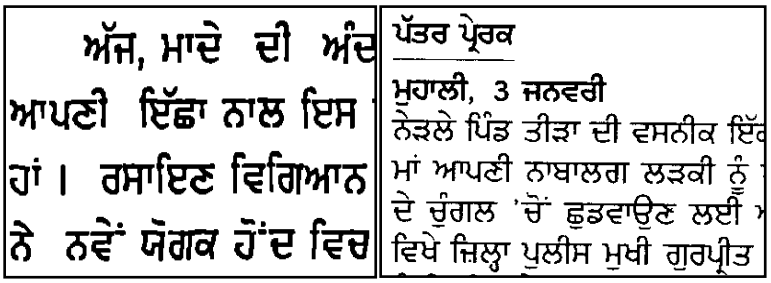
Online handwritten text recognition system used a special surface or digitizer, and a special pen, where a sensor picks up all the movements of pen-tip and also pen-up/ pen-down movements. This is a dynamic approach for recognition.

Off-line handwritten text recognition used written text on any paper or any study material. The text is scanned for recognition. But in both cases writing style, writing method, font style, etc. make it difficult to recognize.

Another way of OCR is machine printed text, which is easier to recognize. The text is printed on pages and these pages are scanned by a good quality scanner with different resolutions or at the same resolutions.

2. Database collection:

A good quality scanners are used to scan the text. A lot of work is done for machine printed text in many languages in India. Gurmukhi is one of the 22 popular languages of India [2, 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.
Existing Gurmukhi OCR is working on character level. For the enhancement of existing OCR, word level images are taken from different sources. A scanner is used to transfer the printed text into computer system in digital form. In this paper, images are scanned at 300 dpi. Some scanned paper images are:



When scanning is completed, preprocessing is applied by binarization and noise removal techniques [4]. After preprocessing next step is to segment the pages into word level. Segmentation stage is done in three categories:

- 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 1:

Table 1: Collection of scanned word images

3. Feature extraction:

Feature extraction is the most important step in OCR system. In feature extraction stage, a set of features is finding which helps to define the shape of text uniquely. 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 select a set of features that uniquely classifies the word. For feature extraction stage Gabor filter and DCT feature are used.

A) Discrete Cosine Transform (DCT):

DCT is the member of a family of sinusoidal unitary transforms [3]. It is a very powerful transform to extract features of the images. DCT encodes the significant details or energy or frequency of the image in a few coefficients very efficiently [5]. These transformed coefficients are used as features of the sample image. It calculates the two-dimensional cosine transform an image [6]. The equation of 2D-DCT image has been represented as:

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

Where:

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

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. The advantage of DCT is that it concentrates the image energy in very few coefficients. 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:

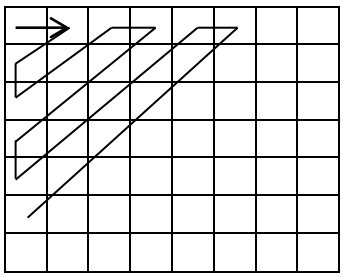


Figure 1. collection of coefficient in zigzag way.

B) Gabor Filter:

Gabor filter is widely applied on the images for feature extraction [7, 8]. 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 [6]. 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 $R_2 = -xsin\phi + ycos\phi$

And λ and ϕ are the wavelength and orientation of plane wave, σ_x and σ_y are the standard deviations of Gaussian envelop along x-axis and y-axis but here $\sigma_x = \sigma_y$. Here the image is resized before feature extraction. For Gabor filter, the image is scaled in 32*32. The x-y plane is rotated by an angle ϕ, which will be result in orientations.the value of ϕ is given by $\phi = \pi(k-1)/m$, where k = 1, 2, …..m. where m denotes the number of orientations. In our case m = 9. Total 189 features are extracted from the image, taking all orientations of the whole image as well as taking each quadrant ad each sub-quadrant. The output 32*32 scaled images for Gabor filter are shown below in table 2,

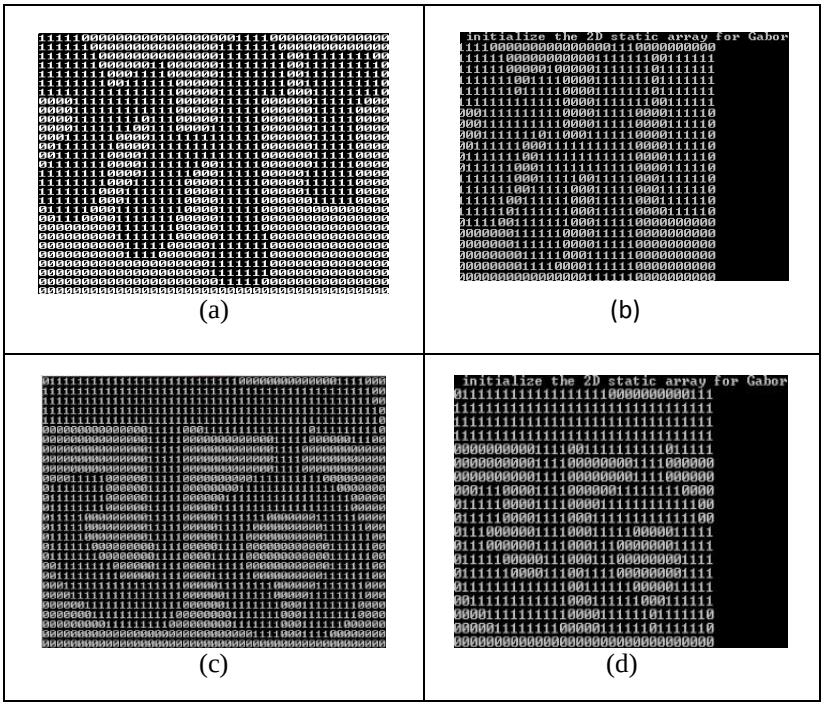


Table 2: (a), (c) original image and (b), (d) scaled 32*32 array for Gabor

4. Conclusion:

Feature extraction is very important and crucial stage in recognition system. This paper concludes two feature extraction techniques, but both are done on monochrome bitmap files. Feature extraction method is the best way to recognize any image, face and text. But it will be enhanced for gray scale images also. And the most important thing: selection of proper feature extraction method because the classifier output depends upon the input features.

5. REFERENCES:

[1] Y. Tawde and M. Kundargi, 2013, "*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.

[2] Kunkari, 2016, "*Optical Character Recognition System for Devanagari Script*", International Journal of Innovative Research in Computer and Communication Engineering", Vol 4, Issue 7, pp 14028- 14033.

[3] Saidas, Rohithram, Sanoj and Manju, 2016, "*Malayalam Charater Recognition using Discrete Cosine Transform*", International Journal of Engineering and Computer Science, Vol 5, Issue 2, pp 15741-15743.

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

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

[6] Singh and Lehal, 2014, " *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.

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

[8] Trier and Jain, 1996, "*Feature Extraction Methods for Character Recognition : A Survey*", Pattern Recognition, Vol 29, No. 4, pp 641 – 662.

Author Biography

Sapna Dhiman: Working as an Assistant Professor of Computer Science Department in M. M. Modi College, Patiala since 2008. Qualification: MCA, M. Phil (Computer Science), Pursuing Ph. D.