IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 12, Issue 5 (Jul. - Aug. 2013), PP 97-102
www.iosrjournals.org

Script Identification for printed document images at text-line level using DCT and PCA

Monali Jindal[1], Dr. Naveen Hemrajani[2]
[1](Computer Science and Engineering Department, Suresh Gyan Vihar University, Jagatpura, Jaipur, India)
[2](Computer Science and Engineering Department, Suresh Gyan Vihar University, Jagatpura, Jaipur, India)

Abstract : *The progress of information technology and the wide reach of the Internet are drastically changing all fields of activity in modern days. As a result, a very large number of people would be required to interact more frequently with computer systems. To develop the human–machine interaction more effective in such situations, it is desirable to have systems capable of handling inputs in a variety of forms such as printed/handwritten paper documents. In a multi-lingual country like India, where more than 22 official languages and 12 different scripts are used for these languages. it is an utmost essential & complicated for designing an OCR system and it became more difficult if the document consist of multiple languages so for an automated multilingual environment such document processing systems relying on OCR would clearly need to identify the script type of the document files, so that specific tool of OCR can be selected.*

In this paper, a script identification approach for Indian scripts is proposed at text-line level. It is a Visual appearance-based script recognition method. The recognition is based upon features extracted using Discrete Cosine Transform (DCT) and Principal Component Analysis (PCA) algorithm and for further extraction we use Modified-KNN. The proposed method is tested on printed document images in 11 major Indian languages, 95% recognition accuracy is obtained.

Keywords: *PCA,DCT,Modified- KNN,OCR*

I. Introduction

The progress of information technology and Internet has drastically changed all fields of activity in modern days. As a result, very large gathering is required to interact more frequently with computer systems. The image of the written text may be sensed "off line" from a piece of paper by optical scanning (optical character recognition).

1.1 Optical Character Recognition: One interesting and challenging field of research in pattern recognition is Optical Character Recognition (OCR).

1.2 OCR in Multi Script Environment: A document containing text information in more than one script is called a multi-script document.

1.3 Automatic Script Identification: Automatic script identification is crucial to meet the growing demand for electronic processing of volumes of documents written in many different scripts.

1.4 Script Recognition Methodologies: Script identification relies on the fact that each script has unique spatial distribution and visual attributes that make it possible to distinguish it from other scripts.

1.5 Script Recognition & Indic Script: A natural way to identify the script in which a document is written may be on the basis of its visual appearance as seen at a glance by a casual observer without really analyzing the character patterns in the document. A survey of offline cursive script word recognition is presented in [1]. The approaches are classified into three categories: segmentation-free methods; segmentation-based methods and the perception-oriented approach. [2]. Chain code based representation and manipulation of hand written images is reported in [3][4]

II. Discrete Cosine Transform (Dct)

The discrete cosine transform (DCT) concentrates energy into lower order coefficients. A DCT shows a bounded sequence of the data points in terms of a sum of cosine functions oscillating at different frequencies. The use of the cosine functions instead of the sine functions is essential for the compression techniques as cosine functions give much more efficiency.

The DCT divides the image into parts in context to the difference in the image visual quality. DCT reduces the correlations of the image data. After the autocorrelation each transform coefficient can be converted independently without losing compression efficiency. For the images we use 2- dimensional DCT in the form of

matrix$M \times N$. The dct2 function calculates the 2- dimensional cosine transform of an image. It is used in the image and the video compression applications.

III. Principal Component Analysis

Principal component analysis follows a protocol by which we can remove the covariance structure of a set of variables. The direction in which a data varies is seen in particular. If variation in data set is caused due to some natural property or due to random experimental error, then it can be called off as normally distributed. [11] Thus we see a distribution by Hyper-Ellipse (as shown in the diagram shown below).

In the first figure the principal direction in which the data varies is shown by U axis. We can see that they are only non zero because of experimental noise. One of the major example where PCA finds is used are:-

Face Recognition has wide usage of Principal component analysis, primarily for reducing the number of variables.

EXPLANATION:- Let's consider the 2D image where we have an input image and we want to compare that image with a set of data base images to see the best fit[11].

HYPOTHETICAL SITUATION is that we consider that the image are all of same resolution and framed correspondingly (Face is appearing at the same place and same scale).

Typically in 2D face images the pixels are not usually in correspondence which is to say a given pixel[xi,yi] may be part of the nose in one image, whereas the same [xi,yi] can be part of his in the other image.[11]

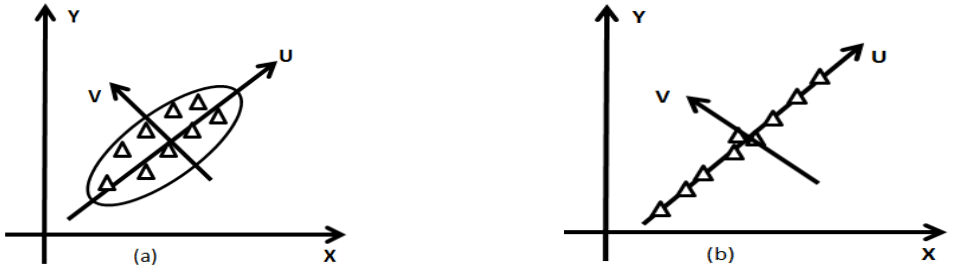

Figure 1: (a) Data Representation Before PCA (b) Data Representation After PCA

IV. Modified K-Nearest Neighbour (KNN) Classifier

K-Nearest Neighbour is also called as non-parametric classifier. Here probability of later probability is calculated from the frequency of nearest neighbour of the pattern which is not known to us. The important concept behind K-nearest neighbour classification is that similar observation is categorised to similar cases. A KNN classifier is taken into consideration for recognition proposes in which normal distance is and not the very common Euclidean distance as normal distance is simple and fast.

This can be shown mathematically as

Training Feature Vector $V_{Train} = [X_1, X_2, X_3, \dots, X_n]$

Where $X_i = [x_1, x_2, x_3, \dots, x_n]$

Test feature vector $V_{Test} = Y = [y_1, y_2, y_3, \dots, y_n]$

For Euclidian distance $D_{X_1,Y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$

For normal distance $D_{X_1,Y} = [|x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n|]$

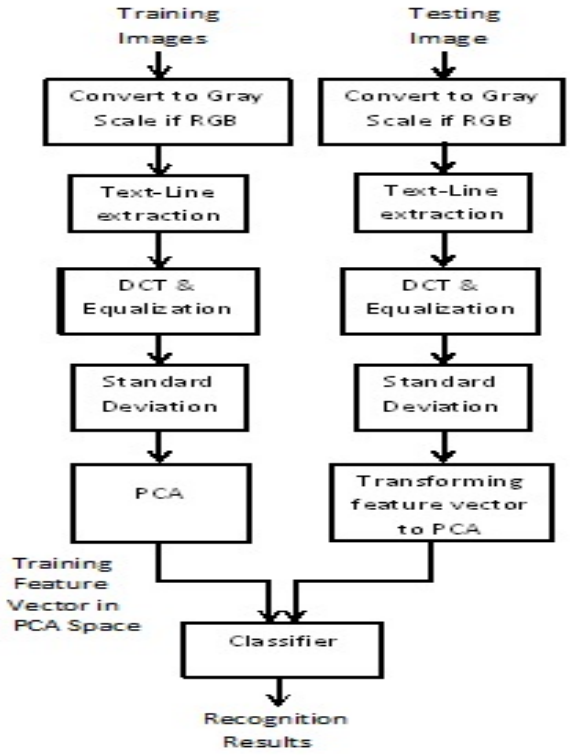

Figure 2: Proposed Script Identification System

V. Literature Review

5.1 STRUCTURE-Based Script Recognition:

Script classes differ from one another in terms of Stroke structure and connections, and the associated writing style. One of the methods of script recognition is to extract connected components in a document [5][6] and later analysing the shapes and structure that we can get to know the internal morphological character of script used.

5.1.1 Text line-wise script identification:

Pal and Chaudhuri have reported majority of the work on Indian language identification. The earliest work we have found on text line method is identification in Indian documents was reported by Pal and Chaudhuri in [7]. This method takes projection profile, statistical and topological features, and stroke features for decision tree-based classification of printed Latin, Urdu, Devnagari and Bengali script-lines. Later, they proposed an automatic system for identification of Latin, Chinese, Arabic, Devnagari and Bengali text lines in printed documents [8].

Based on all these structural characteristics the identification rates obtained were respectively 97.32%, 98.65%, 97.53%, 96.05% and 97.12% for Latin, Chinese, Arabic, Devnagari and Bengali scripts, with an overall accuracy of 97.33%. A more generalized scheme for script-line identification in printed multi-script documents that can classify as many as twelve Indian scripts, viz. Devnagari, Bengali, Latin, Gujrati, Kannada, Kashmiri is available in [9].

5.2 APPEARANCE-Based Script Recognition:

Scripts majorly differ from each other by the shape of individual characters and the way they are grouped in words and sentences. One of the natural ways of identifying a script is on the basis of visual appearence, by giving a glance or by casual observer without analyzing the character patterns.

5.2.1 Page-wise script identification methods:

Visual appearance is mostly linked with texture; block of text corresponding to each script class creates a distinct pattern. In the first step of this method, a uniform text-block on which texture analysis can be performed is produced from the input document image via method given in [12][13]. Texture features are then extracted from the text-block using a 16-channel Gabor filter with channels at a fixed radial frequency of 16 Cycles/sec and at sixteen equally spaced orientations.

The average response of all channels provides a characteristic measure for script that is robust to noise but rotation dependent.

5.2.2 Script identification at paragraph and text-block level:

In the year 2006 G.D. Joshi, S. Garg, and J. Sivaswamy[10], proposed a technique using log-Gabor alters and then classified to different script classes using a KNN classifier. Script identification in Indian printed documents using oriented local energy features was performed in [7]. Local energy is summarised as the sum of squared responses of a pair of conjugate symmetric Gabor filters.

The distribution of energy differs across differently oriented channels of a Gabor filter differs from one script to other. Following figure illustrates how different Indian scripts are classified using these features in two levels of hierarchy.

VI. Problem Statement & Work Done

In a multi-script multi-lingual country like India (India has 18 major regional languages derived from 12 different scripts), a document page like bus reservation forms, question papers, language translation books and money-order forms may contain text lines in more than one script/language forms..

In order to reach a larger cross section of people, it is necessary that a document should be composed of text contents in different languages.

VII. Experimental Results And Analysis

Figure .3: Example of text-line image of each Language

LANGUAGE	IMAGE RECOGNIZED CORRECTLY	IMAGE WORNGLY RECOGNIZED AS	RECOGNITION
Bangla	18/20	Hindi (2)	90%
English	20/20		100%
Gujrati	20/20		100%
Hindi	20/20		100%
Kannada	19/20	Hindi (1)	95%
Malayalam	18/20	Bangla (1), Orriya(1)	90%
Orriya	20/20		100%
Punjabi	19/20	Telegu (1)	95%
Tamil	19/20	Malayalam (1)	95%
Telegu	16/20	Punjabi (2), Kannada (1), Bangla (1)	80%
Urdu	20/20		100%
Total	209/220		95%

Table1:- Recognition results for 20 samples each of eleven languages

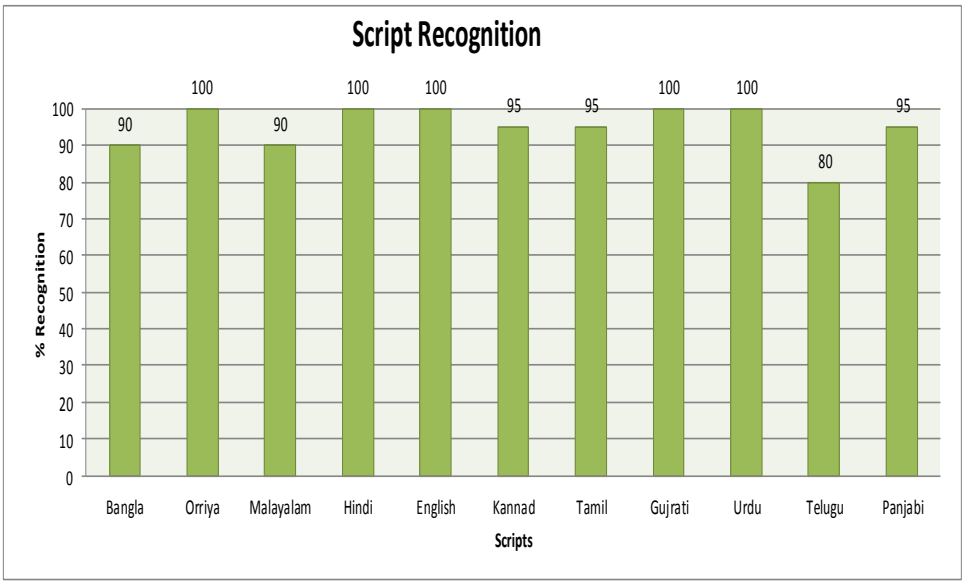

Table 2:-Bar Graph representing 20 samples each of eleven languages

Languages	Bangla	English	Gujrati	Hindi	Kannada	Malayalam	Orriya	Punjabi	Tamil	Telugu	Urdu
Bangla	18			2							
English		20									
Gujrati			20								
Hindi				20							
Kannada				1	19						
Malayalam	1					18	1				
Orriya							20				
Punjabi								19		1	
Tamil						1			19		
Telegu	1				1			2		16	
Urdu											20

Misclassified Correctly Classified

Table 3:-Result analysis for 20 samples each of eleven languages

VIII. Conclusion

Our country acknowledges a wide variety of script, an environment with multiple languages are used, importance arises to clarify the different scripts in terms of writing them and documenting them with authentic character recognition mechanism so as to use document analysis algorithm which be provided a set standard to be used. In this research paper, an approach of script identification has been nominated and specified for all Indian scripts, proposed at text-line level. We call it as Visual appearance-based script recognition method. This recognition is based on feature that is taken from Discrete Cosine Transform (DCT) and Principal Component Analysis (PCA) algorithm. The method that has been proposed in this paper is tested on printed document images which covers 11 major Indian languages that are widely used, and with an accuracy rate of providing 95% recognition thus answering to the queries of languages and scripts in its own accuracy mechanism.

References

[1] Ta Steinherz, Ehud Rivlin, Nathan Intrator, Offline cursive script word recognition: a survey, International journal on Document Analysis and Recognition, IJDAR (1999) 2: 90-110.

[2] Debashis Ghosh, Tulika Dube, and Adamane P. Shivaprasad, Script Recognition—A Review, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 32, No. 12, December 2010

[3] S Madhvanath, G Kim, V Govindaraju, .Chaincode Contour Processing for Handwritten Word Recognition. IEEE transactions on pattern analysis and machine intelligence, 1996, Vol 21 , No 9, pg 928 . 932.

[4] U. Pal and B. B. Chaudhuri, Indian script character recognition: a survey Pattern Recognition, Volume 37, Issue 9, September 2004, Pages 1887-1899.

[5] F. Coulmas, The Blackwell Encyclopedia of Writing Systems, Blackwell Publishers, Oxford, 1996.

[6] C. Ronse and P.A. Devijver, Connected Components in Binary Images: The Detection Problem, John Wiley & Sons, New York, 1984.

[7] U. Pal and B.B. Chaudhuri, "Script Line Separation from Indian Multi-script Documents," Proc. Int'l Conf. Document Analysis & Recognition, Bangalore, pp. 406-409, Sep. 1999.

[8] U. Pal and B.B. Chaudhuri, "Identification of Different Script Lines from Multi-script Documents," Image & Vision Computing, vol. 20, no. 13-14, pp. 945-954, Dec. 2002.

[9] U. Pal, S. Sinha, and B.B. Chaudhuri, "Multi-script Line Identification from Indian Documents," Proc. Int'l Conf. Document Analysis & Recognition, Edinburgh, pp. 880-884, Aug. 2003.

[10] In 2006 G.D. Joshi, S. Garg, and J. Sivaswamy, "Script and Nature Differentiation for Arabic and Latin Text Images," Proc. Int'l Workshop Frontiers in Handwriting Recognition, Niagra, pp. 309-313, Aug. 2002.

[11] DOC493: Intelligent Data Analysis and Probabilistic Interference Lecture 15.

[12] D. Ghosh and A.P. Shivaprasad, "An Analytic Approach for Generation of Artificial Handprinted Character Database from Given Generative Models," Pattern Recognition, vol. 32, no. 6, pp. 907-920, Jun. 1999.

[13] C.L. Tan, P.Y. Leong, and S. He, "Language Identification in Multilingual Documents," Proc. Int'l Symp. Intelligent Multimedia & Distance Education, Baden-Baden, pp. 59-64, Aug. 1999.