ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, January 2015

Script Recognition using GLCM and DWT Features

Vijayalaxmi.M.B [1], B.V.Dhandra [2]

Department of P.G.Studies and Research in Computer Science, Gulbarga University, Gulbarga, Karnataka, India [1, 2]

Abstract: In this paper a method is proposed for identification of Roman, Devanagari, Kannada, Tamil, Telugu and Malayalam scripts at text block level using features of Correlation property of Gray Level Co-occurrence Matrix (GLCM) and multi resolutionality of Discrete Wavelet Transform (DWT) of input handwritten document text blocks. The two-dimensional DWT extracts spatial features and Correlation of GLCM is used to extract texture features. Typically it can be observed that the patterns of any handwritten text block encompass spatial texture primitives. Therefore, the primary aim of this paper is to show the efficiency of DWT and Correlation of GLCM in describing the handwritten text blocks of six Indian scripts. Exhaustive experimentations were conducted on a dataset of 100 text blocks of each script, with bi-script and tri-script combinations of six scripts and script recognition is carried out using three classifiers namely nearest neighbor (NN), LDA and SVM. Using SVM classifier average script classification accuracy achieved in case of bi-script and tri-script combinations are 96.4333% and 93.9833% respectively.

Keywords: Bilingual, Trilingual, Script Recognition, Discrete Wavelet Transform, Correlation of Gray Level Co-occurrence Matrix, texture features, text block level, Nearest Neighbor, Linear Discriminant Analysis, support vector machine classifier.

I. INTRODUCTION

With the growth in information technology capturing, storing and processing of multimedia data is also increasing. Not only printed document processing but also the handwritten document processing has become an inherent part of office automation process. Automatic script and language identification facilitates to read and process the multi-script documents for various applications such as indexing, retrieval of text etc and is an important pre-processing step to optical character recognition. The problem of script identification can be addressed for bi-scripts, tri-scripts and multi-scripts documents.

Automatic handwritten script identification can be classified as: Local and Global approaches. The local approaches employ morphological, water reservoir principle, cavities, corner points, end point connectivity, top and bottom profiles based features used at word or character level and are based on connected components. Basically local approaches are sensitive to noise, improper segmentation, broken characters, and are slower in computation and poor in performance. On the other hand, global approaches involve analysis of large images or the regions (blocks) consisting of two or more text lines, hence segmentation at line, word and character level is not necessary. So script classification task is simple and faster by using global approaches as compared to local approaches. A handwritten text block image contain textural and spatial relationships among the pixels of the image which of course plays a significant role in texture analysis. These observations motivated us to present a generalized global method based on gray level co-occurrence matrix and discrete wavelet transform to address the problem of script identification at text block level.

II. REVIEW OF LITERATURE

Over the last three decades, besides the work on printed text, few works are reported on handwritten text script identification of Indic scripts. Most of the works were focussed on either local or global or combination of local and global approaches of script identification. Guru et al. [9] have given a brief overview and analysis of offline handwritten script identification. Ghosh et al. [8] have given an overview of the different script identification methodologies under each of structure-based and visual-appearance-based categories.

Sharmila et al. [1] have developed a tool for the identification of English, Hindi, Kannada, Tamil, Telugu, Malayalam printed scripts irrespective of their font styles and sizes at word level. The shape, density and transition features were used to perform the nine zone segmentation over the characters. Then script was determined by using rule based classifiers containing set of classification rules which were raised from the zones. The recognition was 89.8%, 92.1%, 86.2%, 97.8%, 89.3% and 86.1% for English, Hindi, Kannada, Tamil, Telugu and Malayalam words respectively. Dhandra et al. [2] have used 13 spatial spread features extracted from morphological filters and classified three handwritten Indian scripts namely English, Devanagari and Urdu based on block level and line level. Using KNN classifier with five fold cross validation an average recognition accuracy of 99.2% for bi-script at text line level and 88.6% for tri-script at block level was achieved. Hangarge et al. [3] have considered automatic handwritten script identification at block level as a texture classification problem. The Gabor filters were used to extract oriented energy features of size 24. The KNN classifier with two fold cross validation gave average tri-script classification accuracy of 91.99 %. Two different methods were used by Hangarge et al. [4] to capture directional edge information. One method by performing

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, January 2015

1D-DCT along left and right diagonals of an image and another by decomposing 2D-DCT coefficients in left and right diagonals. The mean and standard deviations of left and right diagonals of DCT coefficients were computed and considering 9000 word images belonging to six different scripts for validation with linear discriminant analysis (LDA) and K-nearest neighbor (K-NN) classification of the words was performed. At biscripts, triscripts and multiscripts cases respective identification accuracies of 96.95%, 96.42% and 85.77% were achieved. Dhandra et al. [5] extracted Curvelet based features using Discrete Curvelet Transform, nearest neighbor (NN) classification was performed for biscripts and triscripts at block level and obtained average identification accuracies of 94.19% and 90.07% respectively for blocks belonging to six different scripts Roman, Devanagari, Kannada, Tamil, Telugu and Malayalam. Rajput et al. [6] used DCT and Wavelets of Daubechies Family based features and achieved the recognition accuracy of 96.4% using nearest neighbor classifier.

Obaidullah et al. [7] considered six Indian scripts Bangla, English, Devanagari, Urdu, Oriya, Malayalam for script identification. Using some Abstract/Mathematical features, Structure based features, Script dependent features and series of classifiers overall accuracy of 92.8% was obtained on the test set without rejection. They have used a total of 152 documents which include 32 Bangla, 24 Devnagari, 24 Malayalam, 24 Urdu, 24 Oriya and 24 Roman script documents. Out of which 120 were used for training and the rest were used for testing. Kaushik et al. [10] performed word-wise handwritten script identification from bi-script documents written in Persian and Roman. They computed 12 features based on fractal dimension, position of small component, topology etc. and a set of classifiers were employed for script identification experiments. They tested the scheme on a dataset of 5000 handwritten Persian and English words and obtained 99.20% of script identification rate. Bhardwaj et al. [11] extracted moment features for three handwritten scripts Latin, Devanagari and Arabic. The dataset consisted of 12000 word images in training set and 7942 word images in test set and achieved script identification accuracy of over 97% with three classifiers namely decision tree, k Nearest Neighbor (kNN) and Levenberg Marquardt-Nearest Neighbor (LM-NN). Hochberg et al. [12] considered six different scripts Arabic, Chinese, Cyrillic, Devanagari, Japanese and Roman. Using mean, standard deviation, and skew of features namely relative x-centroid, relative y-centroid, number of white wholes, sphericity, aspect ratio discriminated with 88% accuracy across these 6 scripts and found that classification accuracy was higher for documents without fragmented characters and ruling lines. Dhandra et al. [13] have presented an offline writer identification method using gray level co-occurrence matrix based features for English, Kannada and Hindi handwritten documents written by same writer and obtained writer identification accuracies above 80% for all three cases namely single script, bi-script and tri-script writer identification.

In this paper an attempt is made to propose a generic global method using Correlation of GLCM and spatial multi-resolutionality of DWT of handwritten text blocks to discriminate the text patterns of the scripts.

The paper is organized as follows. The description of data collection and feature extraction are presented in Section III. The proposed algorithm and classifiers used in the proposed algorithm are given in Section IV. The experimental results obtained are presented in Section V, followed by conclusion in Section VI.

III. Data Collection And Feature Extraction

A. Data Collection

The standard database for Indian scripts is not available. So the handwritten documents are collected from different writers of different age groups and professions. The collected documents of Roman, Devanagari, Kannada, Tamil, Telugu and Malayalam are scanned through HP Scanjet G2410 scanner to obtain digitized images. The scanning is performed at 300 dpi resolution. The 100 blocks of each script are segmented from the scanned document images. The size of the text block considered for experimentation is 512x512 pixels. Few sample text block images of six scripts are presented in Fig. 1.

Colored document image is converted to gray scale image, which is further binarized using Otsu's global threshold approach.

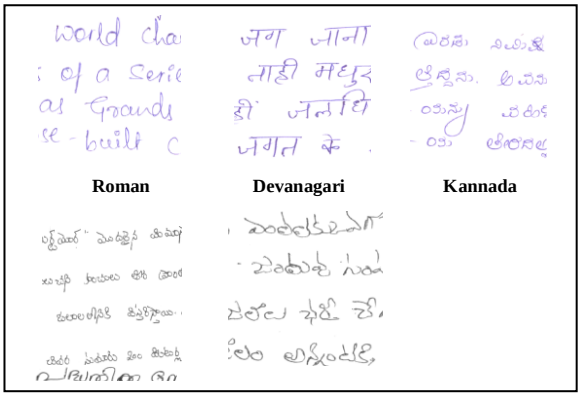

Fig. 1. Sample text blocks in six different scripts

B. Feature Extraction

For feature extraction, we computed correlation of GLCM of input image along 4 directions and 5 distances, DWT with Wavelet family (Coiflet-5) basis function to get the four sub band images namely Approximation (A) and three detail coefficients - Horizontal (H), Vertical (V) and Diagonal (D). The details of feature extraction process are described below.

1) Correlation of Gray Level Co-occurrence Matrix:

A statistical method that considers the spatial relationships of pixels is the Gray-Level Co-occurrence Matrices (GLCM) of the image, also known as the gray-level spatial dependence matrix. We use five distances d = 1, 2, 3, 4, 5 and four directions θ = 0°, 45°, 90°, 135° to construct twenty GLCMs. For each GLCM matrix the common statistical correlation property can be extracted, where $h(x_i,y_j)$ is the (i,j)th entry in the GLCM and is probability

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, January 2015

of occurrence that a pixel with value x_i will be found adjacent to a pixel with value y_j [13].

$$Correlation = \sum_i \sum_j \frac{(x_i - \mu_x)(y_j - \mu_y)\, h(x_i, y_j)}{\sigma_x \sigma_y} \quad (1)$$

where, μ_x, μ_y, σ_x and σ_y are the means and standard deviations of h_x and h_y, and

$$h(x_i) = \sum_j h(x_i, y_j), \qquad h(y_j) = \sum_i h(x_i, y_j),$$

$$\mu_x = \sum_i \sum_j x_i\, h(x_i, y_j), \quad \mu_y = \sum_i \sum_j y_j\, h(x_i, y_j),$$

$$\sigma_x = \sqrt{\sum_i \sum_j (x_i - \mu_x)^2\, h(x_i, y_j)} \text{ and}$$

$$\sigma_y = \sqrt{\sum_i \sum_j (y_j - \mu_y)^2\, h(x_i, y_j)}$$

It is a measure that a pixel is correlated to its neighbour over the whole image. The correlation feature is a measure of gray tone linear dependencies in the image.

2) Discrete Wavelet Transform:

Discrete wavelet transform (DWT) performs sub-band coding on an image in terms of spatial and frequency components and analysis of image from coarse to finer level. The literature on wavelet-based methods continue to be powerful mathematical tools in texture classification problems. The different wavelet transform functions filter out different range of frequencies (i.e. sub bands). Thus, wavelet is a powerful tool, which decomposes the image into low frequency and high frequency sub band images. The wavelet transform breaks an image down into four subsampled images. We have considered only three sub band images namely Approximation (A), Horizontal (H) and Vertical (V) of DWT with Coiflet-5 family.

IV. ALGORITHM AND CLASSIFIERS

During the training phase, features are extracted from the training set. These features are input to classifiers to form a knowledge base that is subsequently used to classify the test images. During test phase, the test image which is to be recognized is processed in a similar way and features are computed as per the algorithm described below.

A. Algorithm Script Recognition

Input: Gray level image of handwritten text block of size

512X512 *pixels.*

Output: Recognized Script

Method: Texture Based Features with NN, SVM and LDA classifiers

Feature vector of size: 23.

Start

Train Phase:

1. Convert colored text block image to gray level image, then gray level image to binary image using Otsu's method. Apply morphological operations to remove noise.

2. For the preprocessed image, obtain 20 GLCMs for 4 directions 0^0, 45^0, 90^0 and 135^0 for five distances d=1, 2, 3, 4, 5. For each GLCM extract the Correlation property, so that 20 features are obtained.

3. Perform Wavelet (Coiflet 5) decomposition for the preprocessed image. And consider only the approximation coefficient (cA), and two detail coefficients horizontal (cH) and vertical (cV) coefficients of the four obtained coefficients. Compute the standard deviation of cA, cH and cV for each frequency band separately. This forms 3 features.

4. Store the computed feature vector of size 23 with script specific labels in the train feature library.

Test Phase:

1. Compute the feature vector of query text block using steps 1, 2, 3 and 4 above.

2. Query text block script is recognized using nearest neighbor, LDA and SVM classifiers.

End.

B. Classifiers

Classifiers used in the proposed method are as follows:

1) Nearest Neighbor (NN):

Basically NN classifier stores the training data *X*. Then finds the minimum distance *d* between training sample *X* and testing sample *Y* using Euclidean distance:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \quad (2)$$

where n is feature vector size.

2) Linear Discriminant Analysis (LDA):

Linear Discriminant Analysis is one of the most commonly used classification technique. It preserves class discriminating information to the higher extent by reducing dimensionality of feature space. It also optimizes separability between the classes by maximizing the ratio of between-class variance to the within class variance. In this paper, LDA is employed on a dataset X=[x1,.....,xi] of dimension N X 23 (N=600) and the sample x_i belongs to one of the class C_i, where i = 1 to 6. Further, the dimension of x_i is mxp, where m = 1 to 100 and p = 1 to 23. Then the classification function is defined as

$$f(X) = Z^T X \quad (3)$$

where Z is the linear projection, and which maximizes between-class scatter

$$S_{between} = \sum_{i=1}^{6} m_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (4)$$

whereas it minimizes the within-class scatter

$$S_{within} = S_1 + S_2 + ... + S_6 = \sum_{i=1}^{6} \sum_{x \in C_i} (X - \mu_i)(X - \mu_i)^T \quad (5)$$

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, January 2015

where μ_i is the mean over class c_i, μ is the mean over all samples, and m_i is the number of samples in class c_i. The classification of a new sample X of class label $\omega \in C_i$ is done based on the nearest neighbor classification rule. For this purpose, the Euclidean distance d of $f(X)$ and the centers $V_i = Z^T \mu_i$ in LDA space are compared.

$$\omega = argmin_{1 \le i \le c}\ d(f(X), V_i) \qquad (6)$$

3) Support Vector Machine (SVM):

The SVM is a the hyper plane classifier with the aim of maximizing a geometrical margin of hyperplane. Subset of training samples closest to margin that determines optimal hyperplane are called support vectors. It involves mapping input vectors X into a high dimensional feature space Z through nonlinear transformation. We have used SVM with Radial Basis Function (RBF) kernel in the proposed method.

V. Experimental Results

The experiments are carried out on 100 text blocks of each script Roman, Devanagari, Kannada, Tamil, Telugu and Malayalam that are segmented from the scanned document images. The size of the text block considered is 512x512 pixels. The proposed method gave outperforming results with nearest neighbor classifier with two-fold cross validation. The average recognition accuracy for bilingual scripts are 94.5667%, 95.9% and 96.4333% using nearest neighbor (NN), LDA and SVM classifiers respectively as shown in Table I and the maximum recognition accuracy is 100% (using LDA and SVM classifier) for Roman-Devanagari, due to dissimilar shapes of the scripts, so the discrimination of the scripts is easy. The minimum accuracy of 82%, 87.5% and 89.5% achieved for Kannada-Telugu using NN, LDA and SVM classifiers respectively and is due to similarity of their character shapes.

The average recognition rate of Kannada-Malayalam, Tamil-Malayalam and that of Malayalam-Telugu are less due to the shape similarity of Kannada, Malayalam, Telugu and Tamil characters. On the other hand average recognition rate for Roman-Devanagari has shown highest accuracy, since Roman and Devanagari scripts are dissimilar.

TABLE I
AVERAGE RECOGNITION ACCURACY OF BILINGUAL SCRIPTS USING 2 FOLD CROSS VALIDATION

Bilingual Script Group	Bilingual Scripts	Recognition accuracy in (%)		
		NN	LDA	SVM
1	R-K	96.5	94.5	95.5
2	R-D	99.5	100.0	100.0
3	R-T	99.0	99.5	100.0
4	R-Tm	91.0	95.5	96.5
5	R-M	97.5	96.0	96.0
6	D-K	95.5	100.0	99.5
7	D-T	98.5	100.0	100.0
8	D-Tm	99.0	99.5	99.5
9	D-M	97.5	97.5	100.0
10	K-T	82.0	87.5	89.5
11	K-Tm	94.5	96.5	97.0
12	K-M	90.0	97.5	91.0
13	Tm-T	98.0	91.0	98.0
14	Tm-M	93.0	90.0	93.0
15	M-T	87.0	93.5	91.0
Average Recognition Accuracy		94.5667	95.9	96.4333

TABLE II
AVERAGE RECOGNITION ACCURACY OF TRILINGUAL SCRIPTS USING 2 FOLD CROSS VALIDATION

Trilingual script group	Trilingual scripts	Recognition accuracy in (%)		
		NN	LDA	SVM
1	RDK	95.3333	96.0	98.6667
2	RDT	98.0	99.6667	100.0
3	RDTm	94.0	97.6667	99.0
4	RDM	95.6667	95.6667	99.3333
5	DKT	87.0	90.6667	92.6667
6	DKTm	94.6667	96.3333	97.6667
7	DKM	86.3333	93.0	92.6667
8	DTTm	96.6667	93.0	98.6667
9	DTM	86.3333	91.0	95.3333
10	DTmM	90.3333	92.6667	95.3333
11	RKT	86.0	85.0	89.6667
12	RKTm	88.0	90.3333	95.6667
13	RKM	86.6667	89.0	94.0
14	RTTm	92.6667	93.6667	94.3333
15	RTM	87.3333	88.3333	96.3333
16	RTmM	89.0	89.6667	92.0
17	KTTm	86.3333	85.3333	87.0
18	KTM	75.6667	80.3333	87.0
19	KTmM	80.0	83.0	87.3333
20	TTmM	83.0	85.0	87.0
Average Recognition Accuracy for all Twenty Combinations		88.95	90.7667	93.9833
Average Recognition Accuracy Excluding KTM Combination		89.6491	91.3157	94.3508

From the Table II, the overall average recognition accuracy of trilingual scripts for all combinations ie., RDK, RDT, RDTm, RDM, DKT, DKTm, DKM, DTTm, DTM, DTmM, RKT, RKTm, RKM, RTTm, RTM, RTmM, KTTm, KTM, KTM and TTmM are 88.95%, 90.7667% and 93.9833% using NN, LDA and SVM classifiers respectively. The maximum recognition accuracy is for Roman, Devanagari, and Telugu and is due to dissimilar shapes of the scripts. Kannada, Telugu and Malayalam have similar shape characters, this leads to the fall in recognition accuracy to 75.6667%, 80.3333% and 87.0% with NN, LDA and SVM classifiers respectively as shown in Table II. In Kannada, Telugu, and Malayalam combination some of the Kannada blocks are misclassified as Malayalam or Telugu, most of the Telugu blocks are misclassified as Kannada or Malayalam due to similarity of Kannada, Telugu and Malayalam scripts. Roman, Kannada and Malayalam, most of the Roman blocks are misclassified as Kannada. This is due to the effect of writing style of native Kannada writer used to write Roman. An attempt shall be made to extract the potential features to discriminate the scripts effectively.

DOI 10.17148/IJARCCE.2015.4157

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, January 2015

Comparative analysis of the proposed method with Hangarge et al.[2] and Dhandra et al.[5] method is shown in Table III.

TABLE III.
COMPARATIVE ANALYSIS OF BLOCK LEVEL TRILINGUAL SCRIPTS IDENTIFICATION

Trilingual Script Group	**Recognition Accuracy in (%) by State of the Art Methods**		**Recognition Accuracy in (%) by Proposed Method using Classifiers (23 features)**		
	Hangarge et al.[2] (24 features)	**Dhandra et al.[5] (20 features)**	**NN**	**LDA**	**SVM**
R-D-K	91.33	93.33	95.33	96.0	98.67
R-D-T	96.00	96.67	98.0	99.67	100.0
R-D-Tm	90.33	95.67	94.0	97.67	99.0
R-D-M	90.33	95.33	95.67	95.67	99.33
Average Rec. Accuracy	91.99	95.25	95.75	97.25	99.25

The proposed method gave 95.75%, 97.25% and 99.25% of recognition rate with RDK, RDT, RDTm, and RDM combinations of scripts using nearest neighbor, LDA and SVM classifiers respectively, where as Hangarge et. al.'s [2] method gave 91.99% and Dhandra et. al.'s [5] method gave 95.25% of recognition rate. The results clearly show that features extracted by using DWT and GLCM yield good results. This enhancing recognition rate of scripts is due to role of mixed features of DWT and GLCM.

VI. CONCLUSION

In this paper, we have presented a technique based on multi-resolution property of DWT and Correlation of GLCM of handwritten text blocks to identify the script at block level. Exhaustive experimentations are carried out on various combinations of scripts and noticed the encouraging performance with the state of the art methods using these mixed features. As segmentation at line, word and character level is not necessary and no connected component analysis is required. It is observed that every script has a distinct textural appearance. Hence the proposed features are used to exploit these properties.

REFERENCES

[1] S.Sharmila, S.Abirami, S.Murukappan, and R. Bhaskaran, "Design and Development of a Script Recognition Tool for Indian Document Images", IJIDCS, Vol. 2 No. 1, 2012.

[2] Mallikarjun Hangarge, Gururaj Mukarambi, and B. V. Dhandra, "South Indian Handwritten Script Identification at Block Level from Trilingual Script Document Based on Gabor Features", In Proc. of Multimedia Processing, Communicating and Computing Applications (Springer:Lecture Notes in Electrical Engineering), Bangalore, pp. no.25-33, 2013.

[3] B.V.Dhandra, and Mallikarjun Hangarge, "Offline Handwritten Script Identification in Document Images", IJCA (0975-8887), Vol. 4, No. 6, July 2010.

[4] M. Hangarge, K.C.Santosh, and P.Rajmohan, "Directional Discrete Cosine Transform for Handwritten Script Identification", In: Proc. of ICDAR 2013, pp. 344-348.

[5] B.V.Dhandra, Vijayalaxmi.M.B, Gururaj Mukarambi, and Mallikarjun Hangarge, "Script Identification using Discrete Curvelet Transforms", International Journal of Computer Applications (0975 – 8887), 2014, pp.16-20.

[6] G. G. Rajput, and Anita H. B, "Handwritten Script Recognition using DCT an Wavelet Features at Block Level", IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition, RTIPPR, 2010.

[7] Sk Md Obaidullah, Supratik Kundu Das, and Kaushik Roy, "A System for Handwritten Script Identification from Indian Document", Journal of Pattern Recognition Research 8 (2013), pp.1-12.

[8] D. Ghosh, T. Dube, and A.P. Shivaprasad, "Script Recognition – A Review", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. XX, No. YY, IEEE, 2009, pp. 2142-2161.

[9] D S Guru, M Ravikumar, and B S Harish, "A Review on Offline Handwritten Script Identification", International Journal of Computer Applications, NCACC, April 2012, pp. 13-16.

[10] Kaushik Roy, Alireza Alaei, and Umapada Pal, "Word-wise Handwritten Persian and Roman Script Identification", In Proc. Twelveth International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 628-633, 2010.

[11] Anurag Bhardwaj, Huaigu Cao, and Venu Govindaraju, "Script Identification of Hand written Images", Proc. of SPIE-IS & T Electronic Imaging, SPIE-IS&T/ Vol. 7247 72470Z-1-6.

[12] Judith Hochberg, Kevin Bowers, Michael Cannon, and Patrick Kelly, "Handwritten Document Image Analysis at Los Alamos: Script, Language, and Writer Identification", 10/1999.

[13] B.V.Dhandra, and Vijayalaxmi.M.B, "Text and Script Independent Writer Identification", International Conference on Contemporary Computing and Informatics, Mysore, Nov. 27-29, 2014.