Entropy Based Texture Features Useful for Automatic Script Identification

M.C. Padma
Dept. of Computer Science & Engg.,
PES College of Engineering,
Mandya-571401, Karnataka, India

P.A.Vijaya
Dept. of Electronics & Communication Engg.,
Malnad College of Engineering,
Hassan-573201, Karnataka, India

***Abstract*—In a multi script environment, a collection of documents printed in different scripts is in practice. For automatic processing of such documents through Optical Character Recognition, it is necessary to identify the script type of the document. In this paper, a novel texture-based approach is presented to identify the script type of the documents printed in three prioritized scripts - Kannada, Hindi and English, prevailed in Karnataka, an Indian state. The document images are decomposed through the Wavelet Packet Decomposition using the Haar basis function up to level two. The texture features are extracted from the sub bands of the wavelet packet decomposition. The Shannon entropy value is computed for the set of sub bands and these entropy values are combined to obtain the texture features. Experimentation conducted involved 1500 text images for learning and 1200 text images for testing. Script classification performance is analyzed using the K-nearest neighbor classifier. The average success rate is found to be 99.33%.**

Keywords-Document Processing Wavelet Packet Tree, Feature Extraction,Script Identification.

I. INTRODUCTION

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 make the man–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. If the computers have to efficiently process the scanned images of printed documents, the techniques need to be more sophisticated. Even though computers are used widely in almost all the fields, undoubtedly paper documents occupy a very important place for a longer period. Also, a large proportion of all kinds of business writing communication exist in physical form for various purposes. For example, to fax a document, to produce a document in the court, etc. Therefore, software to automatically extract, analyze and store information from the existing paper form is very much needed for preservation and access whenever necessary. All these processes go under the title of document image analysis, which has received significance as a major research problem in the modern days.

Script identification is an important problem in the field of document image processing, with its applications to sort document images, as pre processor to select specific OCRs, to search online archives of document images for those containing a particular language, to design a multi-script OCR system and to enable automatic text retrieval based on script type of the underlying document.

Automatic script identification has been a challenging research problem in a multilingual environment over the last few years. All existing works on automatic language identification are classified into either local approach or global approach. Ample work has been reported in literature using local approaches [1-8]. The local features are extracted from the water reservoir principle [1, 3], morphological features [4], profile, cavities, corner points, end point connectivity [7], top and bottom profile based features [5, 8]. In local approaches, the features are extracted from a list of connected components such as line, word and character, which are obtained only after segmenting the underlying document image. So, the success rate of classification depends on the effectiveness of the pre-processing steps namely, accurate Line, Word and Character segmentation. It sounds paradoxical as LWC segmentation can be better performed, only when the script class of the document is known. Even when the script classes are known from the training data, testing requires the performance of LWC segmentation prior to script identification. But, it is difficult to find a common segmentation method that best suits for all the script classes. Due to this limitation, local approaches cannot meet the criterion as a generalized scheme.

In contrast, global approaches employ analysis of regions comprising of at least two text lines and hence fine segmentation of the underlying document into line, word and character, is not necessary. Consequently, the script classification task is simplified and performed faster with the global approach than the local approach. So, global schemes can best suited for a generalized approach to the script identification problem. Satisfactory work has been reported in literature using global approaches [9-18]. Global approaches make use of the texture-based features. These texture features can be extracted from a portion of a text region that may comprise of several text lines.

As most of the Indian states follow three-language policy [1], majority of the documents in an Indian state are printed in three languages such as the regional language, the National language and also English. Accordingly, the documents found in Karnataka an Indian state are generally printed in Kannada the regional language, Hindi the National language and English. A collection of documents printed in these three languages/scripts may be found in some Government and private sectors to facilitate easy access and communication. Few such documents may be found in the department of railways, post offices, banks and such other offices. To automatically process the collection such documents, it is necessary to identify and separate the documents printed in one script so that respective OCR can be employed over them. In this context, this paper proposes a texture-based model for automatic script identification of a collection documents printed in the three scripts - Kannada, Hindi and English.

Texture could be defined in simple form as “repetitive occurrence of the same pattern”. Texture could be defined as something consisting of mutually related elements. Another definition of texture claims that, “an image region has a constant texture if a set of its local properties in that region is constant, slowly changing or approximately periodic”. Texture classification is a fundamental issue in image analysis and computer vision. It has been a focus of research for nearly three decades. Briefly stated, there are a finite number of texture classes C_i, i = 1, 2, 3, n. A number of training samples of each class are available. Based on the information extracted from the training samples, a decision rule is designed, which classifies a given sample of unknown class into one of the n classes [13]. Image texture is defined as a function of the spatial variation in pixel intensities. The texture classification is fundamental to many applications such as automated visual inspection, biomedical image processing, content-based image retrieval and remote sensing. One application of image texture is the recognition of image regions using texture properties. From the literature survey, it is observed that sufficient work has been carried out using texture features. Existing methods on Indian script identification use the texture features extracted from the co-occurrence matrix, wavelet based co-occurrence histogram [12], Gabor filters [20]. Very few works are reported on script identification particularly using wavelet transform based features [12]. In this paper, the features useful for script identification are extracted from the wavelet packets decomposition. As such, no work has been reported that uses the wavelet packet based features for script identification.

The rest of the paper is organized as follows. The Section 2 briefs about the wavelet packet transform. The database constructed for testing the proposed model is presented in Section 3. Section 4 briefs about the necessary preprocessing steps. In Section 5, complete description of the proposed model is explained in detail. The experimental results obtained are presented in section 6. Conclusions are given in section 7.

II. Wavelet Packet Transform (WPT)

Research interest in wavelets and their applications has grown tremendously over the past few years. It has been shown that wavelet-based methods continue to be powerful mathematical tools and offer computational advantage over other methods for texture classification. 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 Continuous Wavelet Transform (CWT) is defined as the sum over all time of the signal multiplied by scaled, shifted versions of the wavelet function ψ:

$$C(scale, position) = \int_{-\infty}^{-\infty} f(t)\psi(scale, position, t)dt \qquad (1)$$

The results of the CWT are many wavelet coefficients C, which are functions of scale and position. The wavelet transform decomposes a signal into a series of shifted and scaled versions of the mother wavelet function. Due to time frequency localization properties, discrete wavelet and wavelet packet transforms have proven to be appropriate starting point for classification tasks. In the 2-D case, the wavelet transform is usually performed by applying a separable filter bank to the image. Typically, a low filter and a band pass filter are used. The convolution with the low pass filter results in the approximation image and the convolutions with the band pass filter in specific directions result in the detail images.

In the simple wavelet decomposition, only the approximation coefficients are split iteratively into a vector of approximation coefficients, and a vector of detail coefficients are split at a coarser scale. That means, for an n-level decomposition, n+1 possible ways of decomposition are obtained as shown in Figure 1. The successive details are never reanalyzed in the case of simple wavelet decomposition.

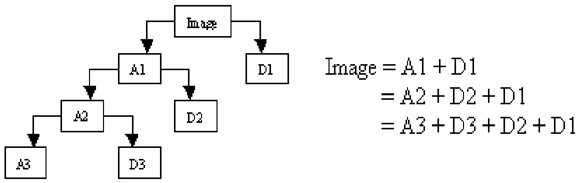

Figure 1. Wavelet Decomposition Tree

The concept of wavelet packets was introduced by Coifman et.al. [23]. In wavelet packets, each detail coefficients vector is also decomposed as in approximation vectors. The recursive splitting of both approximate and detail sub images will produce a binary tree structure as shown in Figure 2.

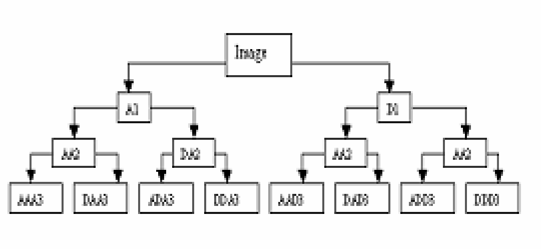

Figure 2. Wavelet Packet Decomposition Tree.

In the case of wavelet, the coefficients of only the approximate sub band are used as the features, whereas in the case of wavelet packets, the coefficients of both approximate and detail sub bands are used as the features. The features derived from a detail images uniquely characterize a texture. The combined transformed co-efficients of the approximate and detail images give efficient features and hence could be used as essential features for texture analysis and classification. A set of decomposed sub bands that yield useful feature values are selected from the wavelet packet tree. The features obtained from the set of sub bands from the transformed images are used for texture classification and are discussed in the Section 5.

III. Data Collection

Standard database of documents of Indian languages is currently not available. In this paper, it is assumed that the input data set contains text blocks of the three scripts - Kannada, Hindi and English, followed in Karnataka, an Indian state. For the proposed model, two sets of database were constructed, one database for learning and the other to test the system. The database of size 500 images for learning and 400 images for testing were used from each of the three scripts. The text blocks 200 images were created using the Microsoft word software for Kannada and English scripts. These text blocks were imported to the Micro Soft Paint program and saved as black and white bitmap (BMP) images. The font type of Times New Roman, Arial, Bookman Old Style and Tahoma were used for English script. The font type of Sirigannada, Vijaya and Kannada Extended were used for Kannada script. The font size of 14, 20 and 26 were used for both Kannada and English scripts. However, the performance is independent of font size. The text blocks of Hindi script was constructed by clipping only text portion of the document downloaded from the Internet. The size of the text block was considered as 600x600 pixels.

One more data set constructed from the scanned document images was used to test the proposed model. The printed documents like newspapers and magazines were scanned through an optical scanner to obtain the document image. The scanner used for obtaining the digitized images is HP Scan Jet 5200c series. The scanning is performed in normal 100% view size at 300 dpi resolution. The image of size 600x600 pixels was considered such that at least 40% of the image contains text region. The test data set constructed from the scanned images involved 200 samples from each of the three scripts.

IV. Preprocessing

Any script identification method used for identifying the script type of a document, requires conditioned image input of the document, which implies that the document should be noise free, skew free and so on. In this paper, the preprocessing techniques such as noise removal and skew correction are not necessary for the manually constructed data sets. However, for the datasets that were constructed from the scanned document images, preprocessing steps such as removal of non-text regions, skew-correction, noise removal and binarization is necessary. In the proposed model, text portion of the document image was separated from the non-text region manually. Skew detection and correction was performed using the existing technique proposed by Shivakumar [22]. Binarization can be described as the process of converting a gray-scale image into one, which contains only two distinct tones, that is black and white. In this work, a global thresholding approach is used to binarize the scanned gray scale images where black pixels having the value 0's correspond to object and white pixels having value 1's correspond to background. It is necessary to thin the document image as the texts may be printed in varying thickness. In this paper, the thinning process is achieved by using the morphological operations. It should be noted that the text block might contain lines with different font sizes and variable spaces between lines, words and characters. Numerals may also appear in the text. It is not necessary to homogenize these parameters. However, it is essential only to ensure that at least 40% of the text block region contains text.

V. The Proposed Model

The proposed model is inspired by a simple observation that every language script defines a finite set of text patterns, each having a distinct visual appearance [13]. Scripts are made up of different shaped patterns to produce different character sets. Individual text patterns of one script are collected together to form meaningful text information in the form of a text word, a text line or a paragraph. The collection of the text patterns of the one script exhibits distinct visual appearance. A uniform block of texts, regardless of the content, may be considered as distinct texture patterns (a block of text as single entity) [13]. This observation implies that one may devise a suitable texture classification algorithm to perform identification of text language. In the proposed model, the texture-based features are extracted from the sub bands of wavelet packet transforms.

A. *Feature Extraction*

In this work, the known input images are decomposed through the Wavelet Packet Decomposition using the Haar (Daubechies 1) basis function to get the four sub band images namely Approximation (A) and the detail - Horizintal (H), Vertical (V) and the Detail (D) coefficients. Through experimentation the Haar basis function up to the level two is found to be best, yielding distinct features and hence Haar with level two is used in this method. This result in a total of 20 sub bands, four sub bands at the first level and sixteen sub bands (four for each sub band) in the next level as shown in Figure 3.

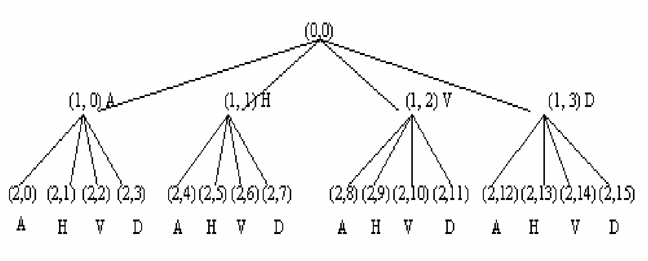

Figure 3. Wavelet Packet Tree up to Level - 2. (A–Approximation, H–Horizontal, V–Vertical and D–Diagonal)

A Shannon entropy value is calculated from each sub bands obtained from the second level wavelet packet tree. Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. Shannon entropy criteria find the information content of signal 'S' using the equation (2)

$$Entropy(S) = -\sum S_i^2 \log(S_i^2) \qquad (2)$$

The sub bands that exhibit the similar type of coefficients from the two levels are selected. For the proposed method, the approximate sub band at level one i.e, (1,0) and the approximate image of (1,0) at level two i.e. (2,0); the horizontal detail sub band (1, 1) at level one, the horizontal sub band (2,1) of (1, 0) at level 2, the approximate (2, 4) and horizontal detail (2, 5) of sub band (1, 1) at level 2 and the vertical detail sub band (1, 2) at level one, the level two vertical sub band (2, 2) of (1, 0), the level two approximate and vertical detail of sub band (2,8) and (2, 10) of (1, 2), are selected to compute the entropy value. Totally, the entropy value is computed from ten sub bands - two from approximate sub band, four from horizontal sub band and four from vertical sub bands. Then, the entropy value of these three set of sub bands (approximate, horizontal and vertical) are added to yield three features. The three entropy based feature values are computed for a training data set of 500 images from each of the three scripts - Kannada, Hindi and English and these features are used as texture features later in the testing stage. The mean value of the texture features obtained from a training data set of 500 images from each of the seven scripts is given in Table 1.

TABLE I. MEAN VALUE OF THE THREE FEATURES - APPROXIMATE, HORIZONTAL AND VERTICAL FOR SEVEN SCRIPT TYPES.

Type of Script	Features values		
	Approximate Feature Value	Horizontal Feature value	Vertical Feature Value
Kannada	62324.35	3413.95	2666.12
Hindi	58475.54	2125.53	2305.79
English	97516.89	2207.93	2990.27

The methodology employed in this paper consists of the following steps.

Algorithm Testing ()

Input: Text portion of the document image containing one script only.

Output: Script type of the test document.

1. Preprocess the input document image.
2. Analyze the test image using 2-d Wavelet Packet Transform with Haar wavelet up to level 2 and obtain the wavelet packet tree.
3. Select the sub bands from the wavelet packet tree as given below:

   Approximation sub bands at the levels: (1,0), (2,0) = (A, AA)

   Horizontal sub bands at the levels: (1, 1), (2, 1), (2, 4), (2, 5) = (H, AH, HA, HH)

   Vertical sub bands at the levels: (1, 2), (2, 2), (2, 8), (2, 10) = (V, AV, VA, VV)
4. Compute the Shannon entropy value from the ten selected sub bands.
5. The entropy value of the sub bands at level one and two are added to get three feature values as given below:

   Approximate feature value = Ent (1,0) + Ent (2,0)

   Horizontal feature value = Ent ((1, 1) + Ent (2, 1) + Ent (2, 4) + Ent (2, 5)

   Vertical feature value = Ent (1, 2) + Ent (2, 2) + Ent (2, 8) + Ent (2, 10)

   where the term Ent refers to Entropy values at the given level.
6. Compute the three feature values from the test image.
7. Classify the script type of the test image by comparing the feature values of the test image with the feature values stored in the knowledge base using *K*-nearest neighbor classifier.

The script type of the test image is classified by comparing the feature values of the test image with the feature values stored in the feature matrix using *K*-nearest neighbor classifier.

B. Classification

In the proposed model, *K* -nearest neighbor classifier is used to classify the test samples. The features are extracted from the test image *X* using the proposed feature extraction algorithm and then compared with corresponding feature values stored in the feature library using the Euclidean distance formula given in equation (3),

$$D(M)=\sqrt{\sum_{j=1}^{N}[f_j(x)-f_j(M)]^2} \tag{3}$$

where *N* is the number of features in the feature vector *f*, $f_j(x)$ represents the *j*th feature of the test sample *X* and $f_j(M)$ represents the *j*th feature of *M*th class in the feature library. Then, the test sample *X* is classified using the *k*-nearest neighbor (*K*-NN) classifier. In the *K* -NN classifier, a test sample is classified by a majority vote of its *k* neighbors, where *k* is a positive integer, typically small. If *K* =1, then the sample is just assigned the class of its nearest neighbor. It is better to choose *K* to be an odd number to avoid tied votes. So, in this method, the *K* -nearest neighbors are determined and the test image is classified as the language type of the majority of these *K*-nearest neighbors. The experiment is conducted for varying number of neighbors like *K* = 3, 5 and 7. The performance of classification was best when the value of *K* = 3.

VI. Experimental Results

The proposed algorithm has been implemented using Shannon entropy computed from the wavelet packet sub bands. The algorithm is tested on two sets of database, each with 200 images. The size of the test images was 600x600 pixels. Elaborate experimentation has been conducted on the images with varying coverage of text. The results of the experiments are given in Table 1. The average classification accuracy of the proposed wavelet based method from both the dataset is 99.33% for full text coverage and 98.8% for partial text coverage. From Table 1, it could be seen that 100% accuracy is obtained for Hindi language for both scanned and manually constructed dataset and also for images covered with 50% texts. The proposed algorithm is implemented using Matlab R2007b. The average time taken to identify the script type of the document is 0.08436 seconds on a Pentium-IV with 1024 MB RAM based machine running at 1.60 GHz. The experimental results demonstrate the effectiveness of the proposed texture features.

TABLE II. Percentage of Recognition of Kannada, Hindi and English scripts.

Script Type	Scanned Dataset		Manually Created Dataset	
	50% text present	Full text covered	50% text present	Full text covered
Kannada	98.1	98.8	98.3	99.4
Hindi	100	100	100	100
English	98.2	98.6	98.2	99.2
Average	**98.77**	**99.13**	**98.83**	**99.53**

VII. Conclusion

In this paper, a global approach for script identification that uses the texture features extracted from the wavelet packet sub bands is presented. The texture features are extracted using the Shannon entropy computed from a set of sub bands. The experimental results demonstrate that the new approach is faster and gives better rate of classification. The performance of the proposed model shows that the global approach could be used to solve a practical problem of automatic script identification.

References

[1] U.Pal, B.B.Choudhuri,: Script Line Separation From Indian Multi-Script Documents, 5th Int. Conference on Document Analysis and Recognition(IEEE Comput. Soc. Press), 406-409, (1999).

[2] U. Pal, S. Sinha and B. B. Chaudhuri: Multi-Script Line identification from Indian Documents, In Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003) 0-7695-1960-1/03 © 2003 IEEE, vol.2, pp.880-884, (2003).

[3] S.Basavaraj Patil and N V Subbareddy: Neural network based system for script identification in Indian documents", Sadhana Vol. 27, Part 1, pp. 83–97. © Printed in India, (2002).

[4] B.V. Dhandra, Mallikarjun Hangarge, Ravindra Hegadi and V.S. Malemath: Word Level Script Identification in Bilingual Documents through Discriminating Features, IEEE - ICSCN 2007, MIT Campus, Anna University, Chennai, India. pp.630-635. (2007).

[5] Lijun Zhou, Yue Lu and Chew Lim Tan: Bangla/English Script Identification Based on Analysis of Connected Component Profiles, in proc. 7th DAS, pp. 243-254, (2006).

[6] M. C. Padma and P.Nagabhushan: Identification and separation of text words of Karnataka, Hindi and English languages through discriminating features, in proc. of Second National Conference on Document Analysis and Recognition, Karnataka, India, pp. 252-260, (2003).

[7] Vipin Gupta, G.N. Rathna, K.R. Ramakrishnan: A Novel Approach to Automatic Identification of Kannada, English and Hindi Words from a Trilingual Document, Int. conf. on Signal and Image Processing, Hubli, pp. 561-566, (2006).

[8] M. C. Padma and P.A.Vijaya: Language Identification of Kannada, Hindi and English Text Words Through Visual Discriminating Features, International Journal of Computational Intelligence Systems (IJCIS), Volume 1, Issue 2, pp. 116-126, (2008).

[9] Santanu Choudhury, Gaurav Harit, Shekar Madnani, R.B. Shet: Identification of Scripts of Indian Languages by Combining Trainable Classifiers, ICVGIP, Dec.20-22, Bangalore, India, (2000).

[10] S. Chaudhury, R. Sheth,: Trainable script identification strategies for Indian languages, In Proc. 5th Int. Conf. on Document Analysis and Recognition (IEEE Comput. Soc. Press), pp. 657–660, 1999.

[11] Gopal Datt Joshi, Saurabh Garg and Jayanthi Sivaswamy: Script Identification from Indian Documents, LNCS 3872, pp. 255-267, DAS (2006).

[12] Hiremath P S and S Shivashankar: Wavelet Based Co-occurrence Histogram Features for Texture Classification with an Application to Script Identification in a Document Image, Pattern Recognition Letters 29, 2008, pp 1182-1189.

[13] T.N.Tan: Rotation Invariant Texture Features and their use in Automatic Script Identification, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 7, pp. 751-756, (1998).

[14] Andrew Busch, Wageeh W. Boles and Sridha Sridharan: Texture for Script Identification, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 11, pp. 1720-1732, Nov. 2005.

[15] A. L. Spitz: Determination of script and language content of document images, IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 19, No.3, pp. 235–245, 1997.

[16] S. L. Wood, X. Yao, K. Krishnamurthy and L. Dang: Language identification for printed text independent of segmentation, Proc. Int. Conf. on Image Processing, pp. 428–431, 0-8186-7310-9/95, 1995 IEEE.

[17] J. Hochberg, L. Kerns, P. Kelly and T. Thomas: Automatic script identification from images using cluster based templates, IEEE Trans. Pattern Anal. Machine Intell. Vol. 19, No. 2, pp. 176–181, 1997.

[18] G. S. Peake and T. N. Tan: Script and Language Identification from Document Images, Proc. Workshop Document Image Analysis, vol. 1, pp. 10-17, 1997.

[19] A. K. Jain and Y. Zhong: Page Segmentation using Texture Analysis, Pattern Recognition 29, pp743-770, 1996.

[20] Hema. P. Menon: Script identification from Document Images using Gabor Filters, Int. conf. on Signal and Image Processing, Hubli, pp. 592-599, 2006.

[21] Rafael C. Gonzalez, Richard E. Woods and Steven L. Eddins,: Digital Image Processing using MATLAB, Pearson Education, (2004).

[22] Shivakumar, Nagabhushan, Hemanthkumar, Manjunath, 2006: Skew Estimation by Improved Boundary Growing for Text Documents in South Indian Languages, VIVEK- International Journal of Artificial Intelligence, Vol. 16, No. 2, pp 15-21.

[23] Coifman R R, Y. Meyer, and M. V. Wickerhauser: Wavelet analysis and signal processing in Wavelets and Their Applications, M. B. Ruskai, Ed. Boston: Jones and Bartlett, 1992, pp.153-178.

AUTHORS PROFILE

M. C. PADMA received her B.E. degree in Computer Science and Engineering from PES College of Engineering (PESCE), Mandya, University of Mysore, Mysore, India and M. Sc. Tech. by Research degree in Computer Science from the Department of Studies in Computer Science, University of Mysore, Mysore, India. She is an Assistant Professor in the department of Computer Science and Engineering, PESCE, Mandya, India. Currently, she is doing her Ph.D. programme in the Department of Electronics and Communication Engineering, Malnad College of Engineering, Hassan, India. Her research areas are multi script document image processing and script identification.

P. A. VIJAYA received her B.E. degree in Electronics and Communication Engineering from Malnad College of Engineering, Hassan, India. She received her M.E. and Ph.D. degrees from the Department of Computer Science and Automation (CSA), Indian Institute of Science (IISc), Bangalore, India. She is a Professor in the department of Electronics and Communication Engineering, Malnad College of Engineering, Hassan, India. Her research interests are in pattern recognition, image processing, data mining, document image processing and script identification.