2014 Sixth International Conference on Computational Intelligence and Communication Networks

A Texture based approach to Word-level Script Identification from Multi-script Handwritten Documents

Pawan Kumar Singh[1], Aparajita Khan, Ram Sarkar, Mita Nasipuri

Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

{pawansingh.ju, aparajitakhan1107, raamsarkar, mitanasipuri}@gmail.com

[1]Corresponding author: {pawansingh.ju@gmail.com}

***Abstract*— Script identification from handwritten document images is an open document analysis problem especially for multilingual environment like India. To design the Optical Character Recognition (OCR) system for multi-script document pages, it is essential to recognize different scripts prior to employing an OCR engine of a particular script. The present work describes a texture based approach to word-level script identification from five handwritten scripts *namely*, *Malayalam*, *Oriya*, *Tamil*, *Telugu* and *Roman*. A 92-element feature vector has been designed in which 80 features consists of selected coefficients of Discrete Cosine Transform (DCT) and the remaining 12 features have been taken from the Moment invariants. Experimentations are conducted on a database consisting of 1000 word images of each script which are evaluated using multiple classifiers. The Multi Layer Perceptron (MLP) classifier is found to be a best choice for the said purpose which is then applied comprehensively using different cross-validation folds and different epoch sizes. The average success rate for the present technique of word-level handwritten script identification is found to be 93.56% for 5-fold cross validation with epoch size 1000, which is quite encouraging.**

Keywords—Script Identification, Handwritten Documents, Discrete Cosine Transform, Moment invariant, Multiple classifiers.

I. INTRODUCTION

Script is defined as the graphic form of writing system which is used to express the written languages. Languages throughout the world are typeset in many different scripts. A script may be used by only one language or shared by many languages, with slight variations from one language to other. For example, *Devnagari* is used for writing a number of Indian languages like *Hindi*, *Konkani*, *Sanskrit*, *Nepali*, etc., whereas *Assamese* and *Bengali* languages use different variants of the *Bangla* script. India is a multilingual country with 22 constitutionally recognized languages written in 12 major scripts. Besides these, hundreds of other languages are used in India, each one with a number of dialects. The officially recognized languages are *Hindi*, *Bengali*, *Punjabi*, *Marathi*, *Gujarati*, *Oriya*, *Sindhi*, *Assamese*, *Nepali*, *Urdu*, *Sanskrit*, *Tamil*, *Telugu*, *Kannada*, *Malayalam*, *Kashmiri*, *Manipuri*, *Konkani*, *Maithali*, *Santhali*, *Bodo*, and *Dogari*. The 12 major scripts used to write these languages are: *Devnagari*, *Bangla*, *Oriya*, *Gujarati*, *Gurumukhi*, *Tamil*, *Telugu*, *Kannada*, *Malayalam*, *Manipuri*, *Sinhala* and *Urdu*. Of these, *Urdu* is derived from the *Persian* script and is written from right to left. The first 11 scripts are originated from the early *Brahmi* script (300 BC) and are also referred to as *Indic* scripts [1-2]. *Indic* scripts are a logical composition of individual script symbols and follow a common logical structure. This can be referred to as the *script composition grammar* which has no counterpart in any other set of scripts in the world. *Indic* scripts are written syllabically and are usually visually composed in three tiers where constituent symbols in each tier play specific roles in the interpretation of that syllable [1].

Script identification aims to extract information presented in digital documents namely articles, newspapers, magazines and e-books. Automatic script identification is useful in sorting document images, choosing appropriate script-specific OCRs and search online archives of document images containing a particular script. Each script has its own character set which is very different from other scripts. However, in this multilingual and multi-script environment, OCR systems need to be capable of recognizing characters irrespective of the script in which they are written. In general, recognition of characters of different scripts with a single OCR module is difficult. This is because of features necessary for character recognition depend on the structural property, style and nature of writing which generally differ from one script to another. Another option for handling documents in a multi-script environment is to use a pool of OCRs (different OCR for different script) corresponding to different scripts. The characters in an input document can then be recognized reliably by selecting the appropriate OCR system from the assumed pool. However, it requires a priori knowledge of the script in which the document is written. Unfortunately, this information may not be readily available. At the same time, manual identification of the documents' scripts may be monotonous and time consuming. Therefore, it is necessary to identify the script of the document before feeding the document to the corresponding OCR system.

978-1-4799-6929-6/14 $31.00 © 2014 IEEE
DOI 10.1109/CICN.2014.60

Difficulties inherent in recognizing handwritten text due to the large variations in handwriting styles pose huge challenges. Resemblances among different scripts are more feasible for handwritten documents rather than for the printed ones. Individual differences, and even differences in the way that people write at different times, enlarge the inventory of possible word shapes seen in the handwritten documents. Also, problems typically addressed in preprocessing, such as ruling lines, word fragmentation due to low contrast, noise removal, skewness, etc. are common in handwritten documents. Since, the script mostly varies from word to word, and not from character to character, so the identification of the scripts at word-level are more preferable than at character or text-line level.

All existing works on automatic language identification are broadly classified into either local approach or global approach [3]. In local approach, the features are extracted from a list of connected components such as text-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 its immediate pre-processing steps. But, it is difficult to find a common segmentation method that best suits for all the script classes. Due to this limitation, local approaches hardly meet the criterion as a generalized scheme. In contrast, global approaches employ analysis of regions and hence fine segmentation of the underlying document into text-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. In the context of local approaches, S. Wood *et al.* [4] described projection profile method to determine *Roman, Russian, Arabic, Korean* and *Chinese* characters. A. L. Spitz [5] proposed a method for distinguishing between *Asian* and *European* languages by examining the upward concavities of connected components. J. Hochberg *et al.* [6] presented a system to automatically identify the six different scripts, *namely*, *Arabic*, *Chinese*, *Cyrillic*, *Devanagri*, *Japanese* and *Roman.* A set of 5 features were extracted from all the connected components assuming eight-connectedness which was trained and tested using Linear Discriminant Analysis (LDA) classifier. R. Sarkar *et al.* [7] proposed 8 holistic features for word-level script identification from *Bangla* and *Devnagari* handwritten texts mixed with *Roman* script by using MLP classifier. P. K. Singh *et al.* [8-9] reported an intelligent feature based technique for word-level script identification of *Devnagari* script mixed with *Roman* script. A set of 39 distinctive features comprising of 8 topological and 31 convex hull based features had been designed which was trained and tested using MLP classifier.

In comparison to local approaches, relatively few works have been reported in the literature for global approaches which, in general, make use of the texture-based features. G. S. Peake *et al.* [10] reported a method for automatic script and language identification from document images using multiple channel Gabor filters and gray level co-occurrence matrices for 7 languages, viz., *Chinese*, *English*, *Greek*, *Korean*, *Malayalam, Persian* and *Russian*. T. N. Tan [11] developed rotation invariant texture feature extraction method for automatic script identification for six languages, viz., *Chinese*, *Greek, English, Russian*, *Persian* and *Malayalam*. Existing methods on *Indic* script identification use the texture features which include wavelet based co-occurrence histogram [12], Gabor filters [13-14], and wavelet packet based features [15]. Global approaches have practical importance in script based retrieval systems because they are relatively fast and reduce the cost of document handling. So, global schemes can be best suited for a generalized approach to the script identification problem. But unfortunately, only a few attempts were made towards word-level handwritten script identification of Indian documents in the literature. This motivates us to use texture based features for word-level script identification written in five scripts *namely*, *Malayalam*, *Oriya*, *Tamil*, *Telugu* and *Roman.*

II. PROPOSED WORK

The proposed scheme is inspired from the observation that humans are capable of distinguishing different objects just by a simple visual inspection. Script types generally differ from each other by the shape of individual characters, and the way they are grouped into words, etc. This gives different scripts distinctively different visual appearances. Texture could be defined in simple form as "repetitive occurrence of the same pattern" or something consisting of mutually related elements. The proposed script identification work consists of texture based features which are extracted from the handwritten word images written in five different scripts *namely*, *Malayalam*, *Oriya*, *Tamil*, *Telugu* and *Roman*. A combination of selected DCT coefficients and moment invariant features has been designed for the said purpose which is described below in detail.

A. Discrete Cosine Transform

DCT [16] is an invertible linear transform that can express a finite sequence of data points in terms of a sum of cosine functions oscillating at different frequencies. It helps to separate the image into parts (or spectral sub-bands) of differing importance (with respect to the image's visual quality). DCT is similar to the discrete Fourier transform. It transforms the original signal or image from the spatial domain to the frequency domain (see Fig. 1) and it is possible to convert back the transformed signal to the original domain by applying the inverse DCT transform.

After the original signal has been transformed, its DCT coefficients reflect the importance of the frequencies that are present in it. For an image of size M×N pixels, the corresponding DCT coefficients are arranged in the form of an 2-D array of size M×N elements as shown in Fig. 1. The top-left corner element of the coefficient array represents the average gray level value in the input image, also known as the DC-coefficient, and usually carries the most representative information of the original signal. The rest of the coefficients, also known as AC-coefficients represent the weightages of higher and higher frequencies along the zig-zag run shown on the coefficient array in Fig. 1. The right-bottom corner element of the coefficient array corresponds to the highest frequency, and generally represents more detailed or fine information of signal and probably has been caused by noise [16]. Information about the image is generally concentrated

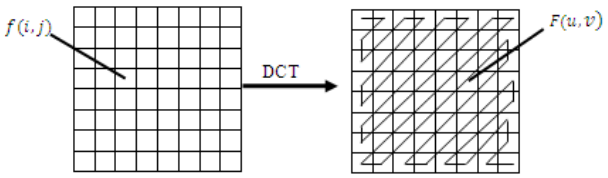

Fig. 1. Illustration of transformation of a function from spatial to frequency domain by DCT [16].

among the coefficients near the top left corner. For an image f(i, j), the 2-dimensional DCT coefficient array F(u, v) is given by:

$$F(u,v) = \sum_{i=0}^{R-1}\sum_{j=0}^{C-1} f(i,j)\cos\left(\frac{\pi u(2i+1)}{2R}\right)\cos\left(\frac{\pi v(2j+1)}{2C}\right) \quad (1)$$

where, R and C are the number of rows and columns of the image array respectively; u and v are the frequency indices along the i and j directions, respectively. A sample word image written in *Tamil* script and its corresponding image after applying DCT and inverse DCT are shown in Fig. 2.

For the present work, the input image is firstly divided into n x n non-overlapping blocks which are known as *grids*. Here, the value of n has been set to 4, as the optimal results have been achieved with this value. The DCT computation is performed on each of the *grids*. Since the pixels in each of the 4 × 4 *grids*, typically have small variations in gray levels, the output of the DCT will result in most of the *grid* energy being stored in the lower spatial frequencies. The corresponding value in the location F(0,0) of the transformed matrix, called the DC coefficient, is the average of all pixel values in the image *grid*. The remaining coefficients are called the AC coefficients and have a frequency coefficient associated with them. Spatial frequency coefficients increase as we move from left to right (horizontally) or from top to bottom (vertically). Low spatial frequencies are clustered in the left top corner. In the present work, the values of the 5 coefficients (*viz*., F(0,0), F (0,1), F (1,0), F (0,2) and F(2,0)) has been computed from each of the *grids*. So, in total, 80 (i.e., 16*5) number of features (F1-F80) has been extracted from each of the word images.

B. *Moment Invariant Features*

Moments are pure statistical measure of pixel distribution around the center of gravity of the image and allow capturing global shape information. They describe numerical quantities at some distance from a reference point or axis. The first significant work considering moments for pattern recognition was performed by M. K. Hu [17]. He derived relative and absolute combinations of moment values that are invariant with respect to scale, position, and orientation based on the theories of invariant algebra which remain invariant under general linear transformations. Geometric moment is defined as the projection of the image intensity function f (x, y) onto the monomial $x^p y^q$ [16]. The (p+q)th order geometric moment M_{pq} of a gray level image f (x, y) is defined as

$$M_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q f(x,y)\,dx\,dy \quad (2)$$

In the case of a digital image of size M×N, the double integral in Eqn. (2) is replaced by a summation which turns into this simplified form as given below:

$$m_{pq} = \sum_{x=1}^{M}\sum_{y=1}^{N} x^p y^q f(x,y) \quad (3)$$

where p, q= 0,1,2,.... are integers.

When f (x, y) changes by translating, rotating or scaling then the image may be positioned such that its *center of mass* (COM) is coincided with the origin of the field of view, i.e. ($\bar{x} = 0$) and ($\bar{y} = 0$), then the moments computed for that object are referred to as *central moment* [18] and it is designated by μ_{pq}. The simplified form of *central moment* of order (p+q) is defined as follows:

$$\mu_{pq} = \sum_{x=1}^{M}\sum_{y=1}^{N} (x-\bar{x})^p (y-\bar{y})^q f(x,y) \quad (4)$$

where, $\bar{x} = \frac{m_{10}}{m_{00}}$ and $\bar{y} = \frac{m_{01}}{m_{00}}$.

The pixel point ($\bar{x}$, $\bar{y}$) is the COM of the image. The *central moments* μ_{pq} computed using the centroid of the image is equivalent to the m_{pq} whose center has been shifted to centroid of the image. Therefore, the *central moments* are invariant to image translations. Scale invariance can be obtained by normalization. The normalized *central moments*, denoted by η_{pq}, are defined as

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \quad (5)$$

where, $\lambda = \frac{(p+q)}{2} + 1$ and (p + q) = 2,3,……

The second order moments $\{\eta_{02}, \eta_{11}, \eta_{20}\}$ are known as the *moments of inertia*, may be used to determine an important image feature called orientation [16]. Here, the feature values F81-F83 have been computed from *moments of inertia* of the word images. In general, the orientation of an image describes how the image lies in the field of view, or the direction of the principal axis. In terms of moments, the orientation of the principal axis, θ, taken as feature value F84, is given by

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2\mu_{11}}{\mu_{20}-\mu_{02}}\right) \quad (6)$$

where, θ is the angle of the principal axis nearest to the x-axis and is in the range $-\pi/4 \le \theta \le \pi/4$. The minimum and maximum distances (r_{min} and r_{max}) between the COM and the boundary of an image are also used as feature descriptors. The ratio r_{max}/r_{min} is called *elongation* or *eccentricity* (F85) and can be defined in terms of *central moments* as follows:

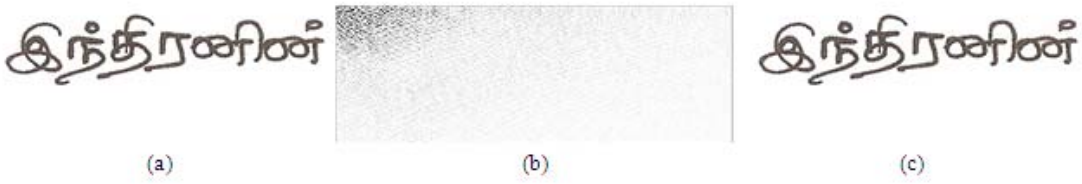

Fig. 2. (a) Sample word image written in *Tamil* script, (b) its corresponding image after applying DCT and (c) reconstruction of the image after applying inverse DCT.

$$e = \frac{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}{\mu_{00}} \qquad (7)$$

The method of moment invariants is derived from algebraic invariants applied to the moment generating function under a rotation transformation. The set of absolute moment invariants consists of a set of non-linear combinations of *central moment* values that remain invariant under rotation. A set of 7 *invariant moments* can be derived based on the normalized *central moments*.

$$\phi_1 = \eta_{20} + \eta_{02} \qquad (8)$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \qquad (9)$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \qquad (10)$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \qquad (11)$$

$$\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \qquad (12)$$

$$\phi_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \qquad (13)$$

$$\phi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \qquad (14)$$

This set of moments is invariant to translation, scale change, mirroring (within a minus sign) and rotation. The 2-D moment invariant gives a total of 7 features (F86-F92) which has been used for the current work.

III. EXPERIMENTAL RESULTS AND DISCUSSION

A total of 5000 words have been collected for the word-level script identification technique. Here, the database for each of 5 scripts *namely*, *Malayalam*, *Oriya*, *Tamil*, *Telugu* and *Roman* contains exactly about 1000 words. The original word images are in gray tone which are digitized at 300 dpi. The word images may contain noisy pixels which have been removed by using Gaussian filter [16]. A well-known Canny Edge Detection algorithm [16] is then applied for smoothing the outer edges of the resultant word images. A total of 3000 words (600 words per script) has been used to train the classifiers whereas the remaining 2000 words (400 words per script) have been used for testing the same purpose. Using a free software tool known as Weka [19], the designed feature set has been individually applied to seven well-known classifiers *namely*, Naïve Bayes, Bayes Net, MLP, Support Vector Machine (SVM), Random Forest, Bagging and MultiClass Classifier. The script identification performances of the present technique on each of these classifiers and their corresponding scores achieved at 95% confidence level are shown in Table II.

From the Table II, it can be seen that MLP classifier shows the best result compared to other classifiers and therefore, MLP classifier has been pushed to its limit to show whether it has the ability to perform better in the current experimental setup or not. For this purpose, we have used 3-fold, 5-fold and 7-fold cross validation schemes with different epoch sizes of MLP classifier. The average identification accuracies achieved after applying different cross validations are shown in Table III. From the table, it is observed that for 5-fold cross validation, MLP produces best result when it is made to iterate 1000 times and the identification accuracy with this set up is found to be 93.56%. The confusion matrix obtained for this best case on the test dataset is shown in Table IV.

The accuracy achieved by the present technique shows convincing results but still some word images have been misclassified. The possible reasons may be due to presence of noise and small words (i.e., words having length of 2-3 characters which produces less discriminating feature values), structural similarity (which in turns causes similarity in the adjacent pixel distribution) in the character set of most of the scripts, presence of abrupt spaces in between characters of a single word image. Sample word images misclassified by the present technique are shown in Fig. 3.

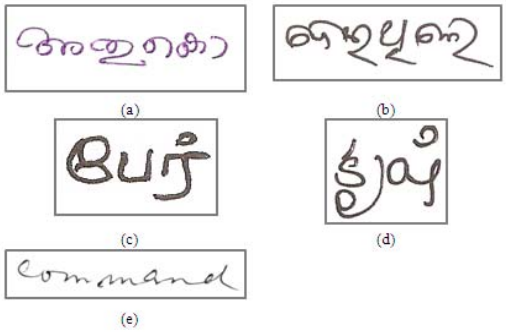

Fig. 3. Sample word images written in *Malayalam*, *Oriya*, *Tamil*, *Telugu* and *Roman* scripts misclassified by the present technique as *Oriya*, *Malayalam*, *Roman*, *Tamil* and *Telugu* scripts respectively.

IV. CONCLUSION

Script identification, a challenging research problem in any multilingual environment, has got attention to the researchers few decades ago. Research in the field of script identification

TABLE I. Success rates of the proposed script identification technique using seven well-known classifiers (best cases are shaded in gray).

	Classifiers						
	Naïve Bayes	**Bayes Net**	**MLP**	**SVM**	**Random Forest**	**Bagging**	**MultiClass Classifier**
Success Rate (%)	81.63	84.73	**91.35**	90.43	89.5	87.46	88.17
95% confidence score (%)	88.5	87.61	**95.78**	93.19	90.03	89.1	90.06

aims at conceiving and establishing an automatic system which would be able to discriminate a certain number of handwritten scripts. As developing a common OCR engine for different scripts is near to impossible, it is necessary to identify the scripts of handwritten text correctly before feeding them to corresponding OCR engine. In this paper, we proposed a texture feature based approach to script identification for some of the *Indic* script documents along with *Roman* script containing handwritten text words. At present, we have used a total of 92 features and the overall accuracy of the system is found to be 93.56%. As the key features used in the technique are mainly texture based, in future, the technique could be applicable as an additional feature for recognizing other scripts in any multi-script environment. More data samples will be collected in future for detailed evaluation of the developed methodology. In a nutshell, the technique could be used as a general word level script identification module for the development of multi-script OCR system.

TABLE II. Recognition accuracies of script identification technique for different folds of cross validation with different epoch sizes of MLP classifier (the best performance is shaded in gray).

	3-fold	**5-fold**	**7-fold**
Epoch size	**Success Rate of MLP classifier (%)**		
500	90.37	92.81	91.45
1000	91.94	**93.56**	92.08
1500	90.72	92.49	91.67

TABLE III. Confusion matrix produced for the best case of the MLP classifier.

Script	***Malayalam***	***Oriya***	***Tamil***	***Telugu***	***Roman***
Malayalam	953	17	5	13	12
Oriya	7	929	18	17	29
Tamil	13	24	924	11	28
Telugu	28	5	12	945	10
Roman	8	25	33	7	927

REFERENCES

[1] H. Scharfe, ''*Kharosti and Brahmi*'', J. Am. Oriental Soc., vol. 122, no. 2, pp. 391-393, 2002.

[2] A.S. Mahmud, ''*Crisis and Need: Information and Communication Technology in Development Initiatives Runs through a Paradox*'', ITU Document WSIS/PC-2/CONTR/17-E, World Summit on Information Society, In: International Telecommunication Union (ITU), Geneva, 2003.

[3] G. D. Joshi, S. Garg, J. Sivaswamy, "*Script Identification from Indian Documents*", In: Lecture Notes in Computer Science: International Workshop Document Analysis Systems, Nelson, LNCS3872, pp. 255-267, Feb. 2006.

[4] S. Wood, X. Yao. K. Krishnamurthi, L. Dang, "*Language identification from for printed text independent of segmentation*", In: Proc. of International Conference on Image Processing, pp. 428-431, 1995.

[5] A. L .Spitz, "*Determination of the script and language content of document images*", In: IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 19, pp.234-245, 1997.

[6] J. Hochberg, K. Bowers, M. Cannon, P. Kelly, "*Script and language identification for handwritten document images*", In: International Journal of Document Analysis and Recognition, pp. 45-52, 1999.

[7] R. Sarkar, N. Das, S. Basu, M. Kundu, M. Nasipuri, D. K. Basu, "*Word level script Identification from Bangla and Devnagari Handwritten texts mixed with Roman scripts*", In: Journal of Computing, vol. 2, issue 2, pp. 103-108, 2010.

[8] P. K. Singh, R. Sarkar, N. Das, S. Basu, M. Nasipuri: "*Identification of Devnagari and Roman script from Multi-script Handwritten documents*", In: Proc. of 5th International Conference on PReMI, LNCS8251, pp. 509-514, 2013.

[9] P. K. Singh, R. Sarkar, N. Das, S. Basu, M. Nasipuri: "*Statistical Comparison of Classifiers for Script Identification from Multi-script Handwritten documents*", In: International Journal of Applied Pattern Recognition (IJAPR), vol. 1, No. 2, pp. 152-172, 2014.

[10] G. S. Peake, T. N. Tan, "*Script and language identification from document images*", In: Proc. of 8th British Mach. Vision Conf., vol. 2, pp. 230-233, Sept.1997.

[11] T. N. Tan, "*Rotation invariant texture features and their use in automatic script identification*", In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp.751-756, 1998.

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

[13] Hema P. Menon, "*Script identification from Document Images using Gabor Filters*", In: International Conference on Signal and Image Processing, Hubli, pp. 592-599, 2006.

[14] P. B. Pati, A. G. Ramakrishnan, "*Word level multi-script identification*", In: Pattern Recognition Letters 29, pp.1218-1229, 2008.

[15] M. C. Padma, P. A. Vijaya, "*Global Approach for Script Identification using Wavelet Packet Based Features*", In: International Journal of Signal Processing, Image Processing and Pattern Recogntion, vol. 20, No. 3, 2010.

[16] R. C. Gonzalez, R. E. Woods, "*Digital Image Processing*", vol. I. Prentice-Hall, India (1992).

[17] M. K. Hu, ''*Visual Pattern Recognition by Moment Invariants*'', In: IRE Transactions on Information Theory, vol. IT-8, pp. 179-187, Feb. 1962.

[18] J. Hochberg, L. Kerns, P. Kelly, T. Thomas, "*Automatic script identification from images using cluster based templates*", In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 2, pp. 176–181, 1997.

[19] www.cs.waikato.ac.nz/ml/weka/documentation.html