A Study of Script Identification Techniques

Rumaan Bashir, Kaiser J. Giri
Islamic University of Science and Technology, Awantipora, J&K
rumaan.bashir@islamicuniversity.edu.in,kaiser.giri@islamicuniversity.edu.in

Abstract
In the recent decades script identification is performed in order to recognize the script(s) of the text present in an image of a document. Since numerous scripts exist in today's world to write the languages therefore a variety of techniques to address this have been developed over time. Some techniques have been applied to the whole image document while as some only target limited areas as a block or even a word. This paper provides a view of most of the techniques which are applied to detect scripts at word-level, line-level, block-level or page-level. Here, a gist of the methods used for script identification is given.

Keywords
Scripts, Script Identification Techniques, Script Identification.

1. INTRODUCTION

Document Image Understanding & Analysis, the area of Computer Science dealing with processing of document images, demands script identification specifically when the possibility of multiple scripts is high and identification is required to proceed to subsequent stages of document processing. Script Identification becomes significantly important when the indexing, retrieval, recognition and understanding are directly going to affect the permissible operations like searching a particular image, sorting images, classifying or categorizing images, selecting suitable script from the given text area or even retrieving document images having text written in a specific script. The fundamental job in script identification is to formulate a method to determine the necessary description of the features present in a script document which almost is sufficient to recognize the document's script for different applications. The methods used for the Identification of Scripts are generally classified into two classes, *local methods & global methods* (Bashir & Quadri, 2014).

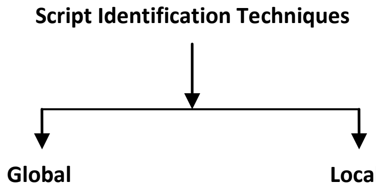

Figure 1: Common Categories of Script Identification Techniques.

In case of local approach, mostly an investigation of connected components is done. Further, here the accomplishment of classification chiefly relies upon segmentation or connected component analysis. Contrary to that, global techniques analyze the regions which comprise more than a single line without further segmentation. However, the local techniques are slower in comparison to the global techniques. Moreover, the script identification is complex because of the actuality that every script possesses unique & distinctive spatial distribution & visual characteristics which makes it different from the other scripts (Bashir & Quadri, 2015).

The techniques of script identification can also be classified on the basis two important parameters, the nature of approach &the features considered (Ghosh et al., 2010)

1. *structure-based techniques*
2. *visual appearance-based techniques.*

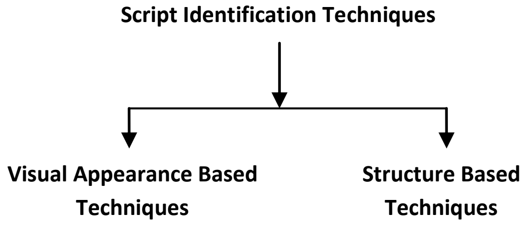
Script Identification Techniques
Visual Appearance Based Techniques
Structure Based Techniques

Figure 2: Types of Common Script Identification Techniques.

2. SCRIPT IDENTIFICATION TECHNIQUES

Script identification techniques have been applied to a variety of scripts all over the world. Different techniques have been proposed with a high degree of outcome and performance. The level at which script identification is applied also has a direct bearing on the classification of the techniques. Various techniques have been developed for script identification which are applied at four major levels to the input document image.

These levels (Ghosh et al., 2010) define the scope of the document image used for recognition and are enumerated as:

1. Word-Level
2. Line-Level
3. Block-Level, and
4. Page-Level.

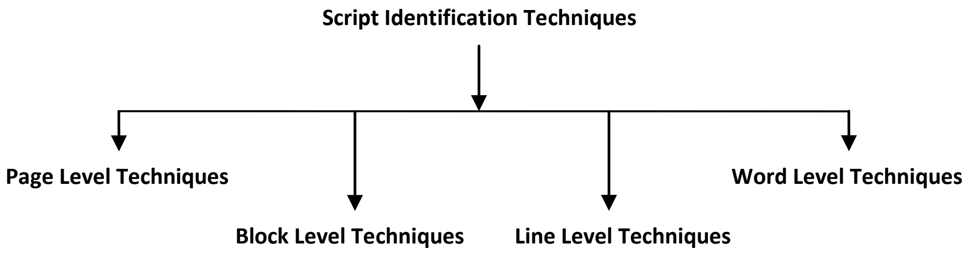

Figure 3: Script Identification Techniques.

The sections 2.1, 2.2, 2.3 and 2.4 discuss the various script identification techniques applied to document images at these levels.

2.1 PAGE-LEVEL SCRIPT IDENTIFICATION TECHNIQUES

A page-level technique is applied to a complete input image of a given document. A range of methods have been presented and implemented at this level including optical density, upward concavities, horizontal projection profiles, connected components, centroids, morphological reconstruction, and other statistical & structural features. Other techniques like texture analysis using texture features, histogram features have also been used for identification at this level. The general scheme is depicted in the following figure.

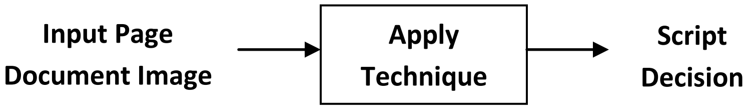

Figure 4: Page-Level Script Identification.

A scheme for identifying the scripts in character structures on the basis of spatial relationships of was presented (Spitz, 1994) for Latin &Han with respect to documents which have machine-printed, wherein the use of structural features like character optical density with the aim to classify individual script-lines in a complete document page was initiated. Similarly, in characters the distribution in the upward concavities vertical for differentiating Latin from Han has been reported (Spitz & Ozaki, 1994). A two-stage classifier was developed (Spitz, 1997) by combining these two aforementioned features.

Some additional features were incorporated (Lee et al., 1996), wherein the script is detected in a printed document using line-wise script identification. Following which, a majority ballot of already determined script-line classification results is performed. Here, top profile, bottom profile, optical density features, upward concavity distribution & character height distribution are the features used. The use of horizontal projections, character density distribution and bounding box size distribution for identifying Cyrillic, Han, Arabic and Latin script in printed format has also been reported (Waked et al., 1998). As compared to the structural features, the aforesaid statistical features have been found to be more robust.

Script identification in typeset document images with the help of statistical features is performed with the help of enclosing structure of connected components & height distributions of connected components and horizontal projection profiles for Latin, Chinese, Japanese or Korean scripts (Lam et al., 1998).A second-level of identification has also been implemented for other scripts with the help of structural features including presence of circles, character complexity, vertical strokes and ellipses.

Script identification in handwritten documents using feature-based method is presented (Hochberg et al., 1999) for differentiating Devanagari, Cyrillic, Arabic, Chinese, Latin and Japanese scripts using standard deviation, mean& skew. Sphericity, relative horizontal centroid, relative vertical centroid, aspect ratio and number of holes of the connected components are the five features which are used for identification in a document page. For classification purposes set of Fisher Linear Discriminants is used.

Identification of script using textual symbols (Hochberg et al., 1997) drawn from documents containing known script by resizing & clustering for template generation for a particular script class has also been presented. These textual symbols comprise fragments of characters, adjoined characters, discrete characters &also whole words. To perform the exact match the symbols, the template symbols are compared to textual symbols extracted from the keyed in document image using Hamming distance. Armenian, Arabic, Burmese, Hebrew, Japanese, Cyrillic, Chinese, Devanagari, Greek, Ethiopic, Korean, Thai and Latin are tested using this technique.

Fractal features based technique was proposed for differentiating printed Japanese, Chinese and Devanagari scripts (Tho & Tang, 2001). For patterns which are extracted from the images of documents, by evaluating fractal signatures the fractal features are generated. The method of evaluating fractal signatures is to calculate the surface area to grey-level function. This grey-level function is equivalent to image of the document.

A font and font-size invariant scheme for detection of scripts in images of printed document for Latin, Devanagari & Urdu based on morphological reconstruction had been proposed (Dhandra et al., 2006). Here, on the image of document in the directions of vertical, horizontal, left & right diagonals with the help of line structuring elements morphological opening &erosion through reconstruction has been performed. Measures of vertical, horizontal, 45°& 135° slanted lines in a document on the basis of the average pixel distributions is performed. Nearest Neighbour classification is used to carry out script identification.

Initial effort to detect scripts in document images which did not analyze the constituent connected component structure was proposed (Wood et al., 1995). Here, with the help of horizontal &vertical projection profiles of images of documents under consideration, the scripts are determined in documents which have been machine printed. This work stressed that in document images projection profiles are enough to distinguish dissimilar scripts.

Visual appearance of documents containing text is seen as a texture and nearly all scripts form a distinct & unique texture pattern. Consequently, script identification can be performed using texture analysis. Texture analysis on the basis of Gabor functions (Tan, 1998) for identification of script in machine printed documents for Latin, Greek, Russian, Chinese, Malayalam &Persian scripts (O'Gorman & Kasturi, 1995) has been presented which is robust to noise. Initially, from the input document page image a uniform text-block is created. By means of sixteen-channel Gabor filter where the channels are at sixteen equally spaced orientations &a fixed radial frequency of 16 Hz, the extraction of texture features from text-block is performed. For rotation invariance, with respect these sixteen channel outputs, Fourier coefficients stand evaluated. With the help of weighted Euclidean distance, the class-representative feature vectors are compared to feature vector produced from text block which are the input for performing classification. For every script and from documents used as a training set, representative feature vectors with respect to a particular script classes have been derived by calculating mean feature vector.

In order to solve the problem of non-uniform character spacing above, (Peake & Tan, 1998) extended the aforementioned effort wherein pre-processing to acquire consistent text-blocks from the document under consideration which is in printed form was performed using simple methods including text line location, spacing normalization, outsized script-line removal, and padding. The GLCM & multi-channel Gabor filter have been implemented individually for feature extraction. GLCMs are being used as a way for attributing texture for a long time now (Haralick et al., 1973). They represent pair-wise joint statistics of pixels in a particular image under consideration. A sixteen-channel filter having 4 frequencies was used at 4 orientations in Gabor filter-based feature extraction. These approaches have been implemented for texture feature extraction in documents with machine printed text which is written in 7dissimilar scripts (Tan, 1998). Later, KNN was implemented to perform script classification.

Due to frequent image filtering, Gabor filter related applications are the high in computational cost. To lessen computation cost, detection of scripts in documents containing machine printed text by means of steerable Gabor filters has been performed (Pan et al., 2005) on Latin, Korean, Chinese & Japanese scripts. It also helps in rotation-invariance. Gabor based functions have shown good results for documents which have been printed by machine but differences in character sizes, character styles, and inter-word &inter-line spacings turn the detection procedure complex once applied directly on handwritten documents. With regards to this, texture-based script detection method was presented (Singhal et al., 2003) wherein pre-processing was performed for tasks like text size normalization, m-connectivity, pruning, thinning &de-noising in reverse order. By means of multi-channel Gabor filter, texture features have been extracted. Latin, Devanagari, Bengali and Telugu scripts are classified with the help of fuzzy classification.

Histogram statistics is one more visual attribute which is employed in a lot of image processing solutions. In an image, spatial distribution of the grey-levels is reflected in histogram statistics. Normalized histogram statistics are used for detecting scripts in Japanese, Cyrillic, Chinese or Latin in typeset documents (Cheng et al., 2006).Here, each script-line of the document image is divided into the following zones — first in between the top-line & x-line called ascender zone, second in between the x-line & baseline called x-zone, and thirdly in between baseline & bottom-line called descender zone. The horizontal projection profiles are calculated with respect to every script-line which provides zone-wise distribution of character pixels in said script-line. Here, it was observed that Cyrillic &Latin characters largely fall in the x-zone having2major peaks positioned on x-line & the baseline where peaks are unequal in Latin and equal in Cyrillic. Chinese characters exhibit a random distribution. They do not form peaks in the profile. In Japanese characters, the average height the profile is having is notably lesser however they also have also exhibit random distribution.

2.2 BLOCK-LEVEL SCRIPT IDENTIFICATION TECHNIQUES

This category of identification of script requires large text blocks as input providing sufficient information to accentuate the characteristics of the script. If the text block is small, performance may get affected. In multi-script document pages, the identification & partition of dissimilar script regions in the image of the document under consideration is essential. The general scheme for block-level identification of scripts is shown in the following figure. Various systems that carry out script detection at the paragraph level are described below.

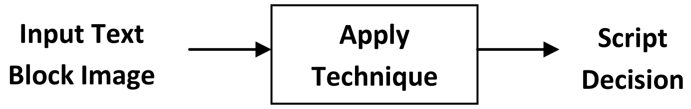

Figure 5: Block-Level Script Identification.

Three different strategies for a printed document image were developed for Latin, Devanagari, Telugu and Malayalam (Chaudhury & Sheth, 1999) to identify the scripts comprising a text block. Here, in the first method the horizontal projection profile's Fourier coefficients are used to describe the script of the block. Consequent to this, classification is performed on the basis of the Euclidean distance in eigenspace. Here, second technique is performed in the text-blocks using Gabor filters through standard deviation & mean on the basis of deriving features from connected components. Third technique is same to the above, except that it uses connected components' distribution of the width-to-height ratio which is present in the document. Afterwards, classification is accomplished using Mahalanobis distance.

A design by means of a Neural network has been prepared for detection with respect to Kannada, Devanagari & Latin scripts (Patil & Subbareddy, 2002). The method comprises two key phases: a modular neural network which follows a feature extractor. Here, morphological operations are used to create a feature vector which corresponds to pixel distributions along specific directions. To structure this modular neural-network,3 independently trained feed-forward neural-networks, one per script are used. On the basis of the network which gives maximum output, input is assigned to that script class. Another, script identification by means of feed-forward neural network without using feature extraction has also been presented (Chi et al., 2003) for Han and Latin text-blocks. Here, the neural network comprises of 4 layers having in input layer forty nine nodes, in hidden layers fifteen & twenty nodes &2 nodes in output layer for depicting two script classes used here.

Latin and Arabic text block identification with respect to handwritten & printed script has been proposed (Kanoun et al., 2002) on the basis of morphological analysis. In addition, at textline & connected component levels geometrical analysis is also applied.

A procedure for Bengali & English script identification (Zhou et al., 2006) applicable in handwritten &machine-printed address-blocks on the images of envelope has been developed. Identification is done through cumulative distance of pixels in bottommost &topmost profiles of connected components. Here, English script exhibits two distance measures which are nearly equal while in the Bengali script image their difference in is apparent.

The use of texture features was used for differentiating printed English &Chinese documents (Jain and Zhong, 1996) proposing a language-independent texture-based page segmentation. This extracts halftone & line-drawing regions and text, automatically from the input greyscale document images. As an addition to this procedure, added segmentation classifies text regions to dissimilar script regions. Here, by means of neural-network training a collection of optimal texture discrimination masks are generated. Afterwards, by convolving trained masks with the image used as input, texture features are derived. Thence, the derived features are employed for classification.

Wavelet energy features (Busch et al., 2005), wavelet scale co-occurrence signatures, wavelets co-occurrence signatures, wavelets log mean deviation features &wavelets log co-occurrence features are used to perform identification of scripts on Latin, Greek, Cyrillic, Hebrew, Han, Japanese, Devanagari and Farsi. Here, 64×64 pixels images of documents which are machine-printed are initially binarized. Following which they are skew corrected &the text-block is normalized (Peake & Tan, 1998).To improve the accuracy & reduce the complexity, Fisher linear Discriminants (FLD) analysis procedure is used. Gaussian Mixture Model classifier with each script class through Gaussian distributions is used to perform classification. With the help of a version of the Expectation Maximization procedure the GMM classifier is trained. A technique on the basis of maximum a posteriori adaptation has been presented as well.

If a particular document image contains one script which is written with only one font, a single model per script class is helpful. On the other hand, various fonts which are normally having generally dissimilar appearance are used for writing a particular script. Due to these variations, probably a model which has been trained for particular set of fonts will not accurately recognize a document image with another font style (Joshi et al., 2006). In view of this, a technique was proposed by means of multiple models per script class (Busch, 2006) to characterize multiple fonts within a single script effectively. After partitioning, linear discriminant analysis is implemented. This is followed by classification with the help of modified MAP-GMM classifier.

Local energy, for Indian printed document script detection (Joshi et al., 2006) has been presented. It is defined as the summation of squared responses of a pair of conjugate symmetric Gabor filters. Efficacy with respect to script classification (Chan & Coghill, 2001) has been demonstrated. With the help of oriented local energy a derived set of descriptors is generated and implemented.

2.3 LINE-LEVEL SCRIPT IDENTIFICATION TECHNIQUES

This category of identification of scripts is applied to individual lines in images of documents. The following figure shows the concept of line-level identification of scripts:

Input Line Image
Apply Technique
Script Decision

Figure 6: Line-Level Script Identification.

The first work was reported for line-level script identification (Pal and Chaudhuri, 1999) for Indian scripts using statistical & topological features, projection profile and stroke features for printed script-lines decision tree-based classification. Automatic scheme for identification of script lines in of Bengali, Arabic, Devanagari, Latin and Chinese in printed documents (Pal and Chaudhuri, 2002) was also presented using 'shirorekha' concept which is used to take apart Devanagari & Bengali script-lines from other script-lines. Subsequently, Devanagari script-lines are differentiated from Bengali by focusing on particular script specific strokes. Likewise, Chinese script-lines are recognized through the presence of four or more vertical runs in characters. At last, Arabic script-lines are differentiated from Latin using statistical & water reservoir features.

Identification of script-line in printed documents has been presented to classify twelve scripts (Pal & Chaudhuri, 2004) using jump discontinuity features, horizontal projection profile, headlines, right &left profiles and water reservoir features.

Connected Component Analysis has been presented (Elgammal & Ismail, 2001) for script-lines using features like the moments, distribution of run-lengths with respect to location-length space and count of peaks in horizontal projection profiles. Here, Arabic script-line horizontal projection profiles show a single peak while English script-line shows two significant peaks. Additionally, the moments in case of Arabic are generally larger than those of English.

Identification of scripts with the help of character component *ngrams* has been patented (Cumbee, 2006). Here, from training documents of a known script segments of characters are extracted. Clustering is performed using K-means clustering. Each script-line is represented by a sequence of numbers which is their corresponding cluster identification number. To find out all the n-grams present, these sequence numbers are scrutinized. Following which, each n-gram is qualified by a weight which corresponds to frequency of occurrence. In identification phase, comparison takes place between character segments in the input script-line and K-means cluster centroids of the known script. Further, n-grams are generated and are compared to n-grams generated in training phase.

2.4 WORD-LEVEL SCRIPT IDENTIFICATION TECHNIQUES

Identification of scripts word-level is performed on individual word images as depicted in the general scheme in the figure below:

Input Word Image
Apply Technique
Script Decision

Figure 7: Word-Level Script Identification.

Many techniques have been applied at word level for identification of scripts. Gabor filter analysis with respect to every word in a bilingual document (Ma & Doermann, 2003; Doermann et al., 2005) is presented where for extracting features characterizing the scripts under consideration. Afterwards, classifier system (two-class) is employed to differentiate between the two scripts. GMM, SVM, weighted Euclidean distance & KNN classifier architectures have also been considered.

Approaches based on visual appearances have been proposed to detect script-words in multi-script documents. Two different approaches in bilingual documents which have been printed (Dhanya & Ramakrishnan, 2002;Dhanya et al., 2002) for identification of scripts at the word-level have been given using three spatial zones & directional energy distribution of words by means of Gabor filters with respect to suitable frequencies & orientations.

An attempt to recognize Tamil & Roman script characters in printed documents using hierarchical design to extract features with the help of zonal occupancy information in addition to structural features is presented in (Dhanya & Ramakraishnan, 2002). Features that are extracted from characters DCT coefficients, DWT coefficients or geometric moments.

A scheme on the basis of Gabor function and performing multi-channel directional filtering (Pati et al., 2004) is applied text area separation &identification of scripts at the word level. This is done by using filter-bank procedure. For separating text/non-text regions this procedure may prove helpful. In addition, Gabor-filter with 4 radial frequencies & orientations is also applied. The extended technique was applied to three - five scripts (Pati & Ramakrishnan, 2006) using filter-bank approach, Gabor filter bank uses 3dissimilar radial frequencies &6dissimilarorientationangles. Classification is performed with Linear discriminant & Nearest neighbour classifier.

3. CONCLUSION

In this paper we have highlighted methods utilized for different studies performed for script identification. A multitude of techniques have been used across various input scopes of the document image. Techniques have been applied at Whole Document level, Text Block level, Script Line Level and even individual Word Level. It is further observed that the same technique may be applied to various levels with successful outcome. Some of these techniques include:

1. Upward Concavities
2. Projection Profiles
3. Texture Analysis
4. Token Based Approach
5. Topological Features
6. Stroke Based Features
7. Structural Features

8. Water Reservoirs
9. Wavelet Transforms
10. GLCM
11. Multichannel Log Gabor filters
12. DCT
13. KNN
14. DCvT and so on.

References

✓ Rumaan Bashir and Quadri, S.M.K., “Entropy based Script Identification of a multilingual Document Image”, IEEE Intl. Conf. Computing for Sustainable Global Development (INDIACom), 2014 Page(s): 19 – 23.

✓ Rumaan Bashir and Quadri, S.M.K., “Density

✓ Debashis Ghosh, Tulika Dube, & Adamane P. Shivprasad, “Script Recognition – A Review”, IEEE, Trans. On PAMI Vol. 32 No. 12 pp 2142-2161 (2010).

✓ A. L. Spitz and M. Ozaki, “Palace: A Multilingual Document Recognition System”, In Proc. IAPR Workshop of Document Analysis Systems pp 16-37 Oct 1994.

✓ A. L. Spitz, “Determination of the script and the language content of document images”, IEEE Transactions on PAMI 19(3) pp 235-245, Mar 1997.

✓ A. L. Spitz, “Multilingual Document Recognition”, In Proc. Intl. Conf. on Electronic Publishing, Document Manipulation, and Typography, pp 193-206, Oct 1994.

✓ D. S. Lee, C. R. Nohl and H. S. Baird, “Language Identification in Complex, Unoriented, and Degraded Document Images”, In Proc. IAPR Workshop of Document Analysis Systems pp 76-98 Oct 1996.

✓ B. Waked, S. Bergler, C. Y. Suen and S. Khoury, “Skew detection , Page segmentation and Script classification of Printed document images”, In Proc. Intl. Conf. on Systems, Man and Cybernetics, Vol 5 pp4470-4475, Oct 1998.

✓ L. Lam, J. Ding and C. Y. Suen, “Differentiating between Oriental and European Scripts by Statistical Features”, Intl. Conf. on Pattern Recognition and Artificial Intelligence Vol. 12 No. 1 pp 63-79, Feb 1998.

✓ J. Hochberg, K. Bowers, M. Cannon and P. Kelly, “Script and Language Identification for Handwritten Document Images”, Intl. Journal of Documents Analysis and Recognition, Vol. 2, Nos. 2-3 pp 45-52 Dec 1999.

✓ Judith Hochberg, Patrick Kelly, Timothy Thomas & Lila Kerns, “Automatic Script identification From Document Images Using Cluster-Based Templates”, IEEE Analysis and machine Intelligence, Vol. 19, No. 2, Feb 1997.

✓ Y. Tho and Y. Y. Tang, “Discrimination of Oriental and Euramerican Scripts Using Fractal Feature”, In Proc. Intl. Conf. Document Analysis and Recognition, pp 1115-1119, Sept 2001.

- B. V. Dhandra, H. Mallikarjun, Ravindra Hegadi and V. S. Malemath, “Word-wise Script Indentification from Bilingual Document Based on Morphological Reconstruction”, IEEE 2006.
- S. L. Wood, X. Yao, K. Krishnamurthi, and L. Dang, “Language Identification for Printed Text Independent Segmentation”, In Proc. Intl. Conf. On Image Procesing, Vol 3, pp 428-431, Oct 1995.
- T. N. Tan, “Rotation Invariant Texture Featues and Their Use in Automatic Script Indentification”, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 20, No. 20, July 1998.
- L. O’Gorman and R. Kasturi, “Document Image Analysis”, IEEE CS Press, 1995.
- G. S. Peake & T. N. Tan, “Script and Language Identification from Document Images”, In Proc. Asian Conf. Computer Vision, pp. 97-104, Jan 1998.
- R. M. Haralick, K. Shanmugam and I. Dinstein, “Textual features for image classification”, IEEE Transations on Systems, Man, and Cybernetics, Vol. 3 No. 6 pp 610-621, Nov 1973.
- W. M. Pan, C. Y. Suen and T. D. Bui, “Script Identification Using Steerable Gabor Filters”, In Proc. 8th Intl. Conf. on Document Analysis and Recognition , 2005
- V. Singhal, N. Navin and D. Ghosh, “Script-Based Classification of Hand-written Text Documents in a Multilingual Environment”, In Proc. of the 13th IEEE International Workshop on Research Issues in Data Engineering: Multi-Lingual Information Management (RIDE-MLIM '03), pp. 47–54, March 2003.
- Juan Cheng, Xijian Ping, Guanwei Zhou and Yang Yang, “Script Identification of Document Image Analysis”, In Proc. IEEE 1st Intl. Conf. on Innovative Computing, Information and Control 2006.
- Santanu Chaudhury and Rabindra Sheth, “Trainable Script Identification Strategies for Indian Languages”, In Proc. of the 5th International Conference on Document Analysis and Recognition (ICDAR '99), pp. 657–660, 1999.
- B. Patil and N. V. Subbareddy, “Neural network based system for script identification in Indian documents,” Sadhana, vol. 27, No 1, pp. 83–97, 2002.
- Zheru Chi, Qing Wang and Wan-Chi Siu, “Hierarchical content classification and script determination for automatic document image processing”, Elsevier Pattern Recognition, 36 (2003), 2483-2500.
- S. Kanoun, A. Ennaji, Y. Lecourtier, and A.M. Alimi, “Script and Nature Differentiation for Arabic and Latin Text Images”, In proc. Intl. Workshop Frontiers in Hnadwriting Recognition, pp 309-313, Aug 2002.
- L. Zhou, Y. Lu and C. L. Tan, “Bangla/English Script Identification Based on Analysis of Connected Component Profiles”, In Proc. Intl. Workshop on Document Analysis Systems, pp 243-254, Feb 2006.

✓ A. K. Jain and Y. Zhong, “Page segmentation using texture analysis”, Pattern Recognition, Vol. 29, No. 5, pp730-770, May 1996.

✓ Busch, Andrew ; Boles, W.W. ; Sridharan, S. , “Texture for script identification”, Pattern Analysis and Machine Intelligence, IEEE Transactions on Volume: 27 , Issue: 11 (2005) , Page(s): 1720 – 1732.

✓ G. D. Joshi, S. Garg and J. Sivaswamy, “Script Identification from Indian Documents”, In Proc. IAPR Intl. Workshop on Document Analysis Systems, pp 255-267, Feb 2006.

✓ A. Busch, “Multi-Font Script Identififcation Using Texture-Based Features”, In Proc. Intl. Conf. Image Analysis and Recognition, pp 844-852, Sept 2006.

✓ Woei Chan, George Coghill, “Text analysis using local energy”, Elsevier Pattern Recognition, 34, (2001) 2523-2532.

✓ U. Pal & B. B. Chaudhari, “Script Line Separation From Indian Multilingual Script Documents”, Proc 5th Intl. Conf, on Document Analysis ad Recognition, IEEE Comp. Society Press pp 406-409 (1999).

✓ U. Pal and B. B. Chaudhuri, “Identification of Different Script Lines from Multi-Scripts Documents”, Image and Vision Computing, Vol. 20, No. 13/14, pp. 945-954, Dec. 2002.

✓ U. Pal & B. B. Chaudhuri, “Indian Script Character Recognition: A survey”, Elsevier Pattern Recognition 37 (2004) 1887-1899.

✓ Ahmed M Elgammal & Mohamed A. Ismail, “Techniques for Language Identification for Hybrid Arabic-English Document Images”, IEEE, *ICDAR, page 1100-1104. IEEE Computer Society,* (*2001*).

✓ C. S. Cumbee, “Method of Identifying Script of Line of Text, US Patent 7020338, Mar. 2006.

✓ H. Ma and D. Doermann, “Gabor Filter Based Multi-Class Classifier for Scanned Document Images”, In Proc. Intl. Conf. Document Analysis and Recognition, pp. 968-972, Aug. 2003.

✓ S. Jaeger, H. Ma and D. Doermann, “Identifying Script on Word-Level with Informational Confidence”, In Proc. Intl. Conf. Document Analysis and Recognition, Vol. 1, pp. 416-420, Aug. 2005.

✓ D. Dhanya and A. G. Ramakrishnan, “Optimal Feature Extraction for Bilingual OCR”, In Proc. IAPR Intl. Conf. Workshop Document Analysis Systems, pp 25-36, Aug 2002.

✓ D. Dhanya and A. G. Ramakrishnan, “Script identification in Printed Bilingual Documents”, In Proc. IAPR Intl. Conf. Workshop Document Analysis Systems, pp 13-24, Aug 2002.

✓ D. Dhanya, A. G. Ramakrishnan, and P.B. Pati, “Script identification in Printed Bilingual Documents”, Sadhana, Vol. 27, No. 1 pp 73-82, Feb 2002.

✓ P. B. Pati, S. Sabari Raju, N. Pati and A. G. Ramakraishnan, “Gabor Filters for Document Analysis in Indian Dilingual Documents”, In Proc. Intl. Conf. Intelligent Sensing nd Infromation Processing, pp 123-126, Jan 2004.

✓ P. B. Pati and A. G. Ramakraishnan, “HVS Inspired System for Script Identification in Indian Multi-Script Documents”, In Proc. Intl. Workshop Document Analysis Systems, pp 380-389, Feb 2006.