Automatic Multi-lingual Script Recognition Application

Waleed Abdel Karim Abu-Ain
wabuain@yahoo.com
King Abdulaziz University, Saudi Arabia

Siti Norul Huda Sheikh Abdullah
snhsabdullah@ukm.edu.my
Center for Cyber Security,
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia

Khairuddin Omar
ko@ukm.edu.my
Center for Artificial Intelligence Technology
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia

Siti Zaharah Abd. Rahman
sitizaharaharab@gmail.com
Center for Cyber Security,
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia

ABSTRACT

Document Image Analysis and Recognition (DIAR) technique is used to recognize text component and translate it into editable format. Scripts are a set of graphical representations used to express a particular writing system as well as subsets belonging to a particular writing system. The writing styles of more than one script family may then be adopted by one language, such as in the cases where the old Malay language (Jawi) adopts the Arabic script while the modern one adopts the Roman script. The seven major scripts used in this research are in handwritten style including Arabic, Devanagari, Hebrew, Thai, Greek, Cyrillic and Korean. Automatic Multi-lingual Script Recognition (AMSR) is one of the main challenges in DIAR domain. Currently, only few attempts have been made for automated script identification of off-line handwritten documents images. Most available AMSR applications only deal with printed documents and script types, and they neglect handwritten and multi-lingual documents. The objective of this study is to propose a multi-lingual AMSR framework. The research methodology consists of a proposed multilingual AMSR framework. The multilingual AMSR framework is tested on Multilingual-HW datasets, which contains more than seven international unconstraint handwritten scripts, using Grey-Level Co-occurrence Matrix and Local Binary Pattern. The average accuracy of both methods is about 97.01% and 85.29% respectively. This proposed multilingual AMSR is hoped to be beneficial to a group of community which requires automatic sorting multi-lingual documents. This research can also be extended to document forensic area or international relations agency to identify unknown native document.

Keywords: Automatic Multi-lingual Script Recognition (AMSR); feature extraction; statistical texture analysis; Grey-Level Co-occurrence Matrix (GLCM); Local Binary Pattern (LBP)

INTRODUCTION

Across multiple science disciplines, Artificial Intelligence (AI) is lauded as one of the most crucial research domains as well as one of the world's most prominent high technology in the current century. The holy grail of AI researches includes designing a machine vision, which is capable of simulating the intelligence of an ideal human. It entails the ability of recognition, reasoning, learning, communication, feeling and much more. Furthermore, AI covers diverse areas of research including robotics, machine learning, language processing, machine vision, intelligence agents, gaming, neural networks and pattern recognition (PR). PR, being one of the most important domains in AI concerns towards classifying the patterns based on their features (X. Tan & Triggs, 2010).

In computer science, machine vision is an essential area of AI in which a machine or a robot is accorded with the ability to understand, recognize, categorize and extract useful information from images. It entails the construction of image features by the description and location of the pattern based on its physical attributes (Ratha & Jain, 1999). The application of computer vision encompasses a variety of fields such as image retrieval (Chen et al., 2009; Kulkami et al., 2014), satellite image classification (Najab 2010, Raj & Sivasathya 2014), medical image diagnosis (Mohamed, 2008; Giger & Pritzker, 2014), biometric recognition (Azizi & Pourreza, 2009; Sanchez et al., 2014), and document images analysis and recognition (Abdullah et al., 2006; Journet et al., 2008; Bayro-Corrochano & Hancock, 2014; Sulaiman, Omar & Nasrudin, 2017; Bataineh et al., 2017).

Document Image Analysis and Recognition (DIAR) mainly addresses the issue of isolating texts from the graphics to enable localisation for recognition. DIAR is granted the central stage in pattern recognition domain which considers the textual analysis as the core part of DIAR. The objective of its undertaking is to recognise the text component and translate it into editable format. The available methods in textual processing include optical script recognition, skewing, text line, and words (Kasturi, O'gorman & Govindaraju, 2002).

The applications of DIAR are ubiquitous in our daily endeavor, as evidenced on several areas of document image enhancement such as binarisation (Tensmeyer & Martinez, 2017), skew correction (Vinod & Niranjan, 2018) and text extraction (Saabni, Asi & El-Sana, 2014). DIAR bears an imperative role in document salvaging tasks namely document image and information retrieval (Ahmed, Al-Khatib & Mahmoud, 2017), old and historical manuscripts (Saabni et al., 2014), digital libraries (Abidi, Siddiqi & Khurshid, 2011), and information extraction from images (Chen et al., 2011). Optical character recognition (Singh et al., 2010), optical font recognition (Lutf et al., 2014), and optical script recognition (Rao, Imanuddin & Harikumar, 2014) are also among the important areas in which recognition and understanding process are deemed necessary.

Based on a survey conducted on the DIAR; there are two categories in which DIAR are frequently applied: recognition and processing applications (Marinai, 2008). Any application revolving around recognition will require the reasoning processes to recognise and extract information from document images. This entails the utilisation of a set of applications such as optical script recognition (OSR), signature verification, optical character recognition (OCR), language identification and writer identification.

On the other hand, any processing applications that deal with pixels are geared towards enhancing the quality of document images as image. These applications are not intended to extract information from the document image but to represent a set of applications such as binarisation, noise removing, filtering, layout analysis, text segmentation and skew correction (Marinai, 2008). Each one of the category may embody the textual or graphical objects in the document images. Graphical application of DIAR focuses on the objects present in document images such as pictures, logos and figures. It involves several state of art techniques including preparation of graphical regions, filling of the graphical regions, graphic localisation and extraction, logo recognition, and finally graphic recognition. DIAR is applied extensively in the field textual enhancement that seeks to extract information related with text recognitions.

Apart from that, Optical Character Recognition (OCR) is also presented as one of the most concerned research field in DIAR. OCR is a tool for converting the textual parts in the image to encoded texts. Several OCR methods are currently available concentrating on both printed (Radwan, Khalil & Abbas, 2017) and handwritten forms (Boufenar, Kerboua & Batouche, 2018; Kamble & Hegadi, 2015). The OCR algorithm manages to yield an acceptable performance for printed form but it is, otherwise, poor for handwritten, requiring further improvement.

Optical script recognition (OSR) technique is another interesting applicatory domain in DIAR. Instead of font type, this technique enables automatic identification of script language in document images, making it a prerequisite for documents containing multiple languages. According to Joshi, Garg and Sivaswamy (2006), OSR lays claim to a number of advantages such as improvement in OCR operation, as well as sorting and searching document.

The objective of this study is to propose an automatic multi-lingual script recognition system. This system is able to classify and sort document images into several international languages such as Arabic, Devanagari, Hebrew, Thai, Greek, Cyrillic, and Korean. This paper is organised into six sections namely introduction, literature review, proposed method, experimental results and analysis, discussion and conclusion.

LITERATURE REVIEW

All pattern recognition (PR) operations, generally undergo the main three stages namely, pre-processing, feature extraction and recognition. According to Marinai (2008), the recognition approaches bear a similarity with the applications of DIAR such as OCR, OSR, OFR and Logo Recognition comprising pre-processing, feature extraction and recognition phases subsequently.

The pre-processing stage is a process of document image preparation before exiting to the next stage. This pre-processing stage may embody different sequential approaches conducted on the document input image. As a result, some applications require the segmentation process to be performed in the pre-processing stage if local feature extraction methods are used whereas it is not required for global feature extraction. In addition, the filtering process is compulsory if the document images are heavily distorted with noises and skewed. For some of DIAR applications, text thinning is required in the pre-processing stage.

The next stage in recognition approach in DIAR and PR is the feature extraction which entails the conversion of the objects in the document images to numerical or discreet information readable by machines. A good feature extraction method is characterised by the ability to extract similar information about similar objects and different information for different objects (Bataineh, Abdullah & Omar, 2012; Zavvar et al., 2016). The feature extraction methods comprise two approaches; local approach (Li et al., 2009) and global approach (Bataineh et al., 2012; Ubul et al., 2017). The local feature extraction is based on segmentation process wherein the text in the document images is split into several parts (character, glostrokes, primitives) to be used for extracting discreet information on features values. The global feature extraction methods, on the other hand, extract features values by analysing the properties of the document images surface from which the statistical or numerical information is acquired.

Finally, recognition is allocated in the last stage in DIAR and PR operation. Therein the extracted information and features of an object from the previous stage is distinguished based on defining characters. The recognition stage, which can take the form of prediction, classification, clustering, is executed via a range of methods such as neural networks, rule base, tress and much more. In fact, the objective of recognition process in DIAR is to identify the class for which the scoped object in the document images belong to Jiang (2009). This is evidenced by the application of OSR in identifying script language of the text in the input image.

Optical character recognition (OCR) is a vital step in the concept of AMSR system in document image analysis and recognition (DIAR) area. Offline OCR system aims to convert text from document images into an editable format. In international environment system, the prior knowledge of the script is crucial and mandatory to channel the document image into the proper OCR system. For this reason, finding an optical script identification (OSR) is a crucial necessity. Moreover, OSR provides a significant information about the document such as content, historical information and the document structure. OSR aims to recognise the document script type automatically and thus, is considered as the main step in any full automatic multi-OCR system. Although most existing OSR methods explore machine-printed documents domain, few have dealt with handwritten documents domain and the unconstraint-handwritten scripts are almost neglected (Ghosh, Dube & Shivaprasad, 2010).

The following sections present an overview of the different local and global approaches that have been proposed to address the problem of OSR of off-line document images in both the machine-printed and handwritten text domains.

MACHINE-PRINTED DOMAIN IN OSR APPLICATION

In document image analysis and recognition (DIAR), Gabor energy feature is an ideal feature extraction method for identifying the scripts of word in a multilingual machine-printed document (Peete & Ramakrishnan, 2008). Gabor energy features encompassing different scales and orientations in different direction filter have been used to localize and extract text-only regions from complex document images (Khaleefah & Nasrudin, 2016). Gabor energy feature was first used in texture analysis application to segment document images into Chinese and Latin script text areas for the script recognition problem by Jain and Zhong (1996). Tan (1998) used a weighted Euclidean distance classifier to compare the rotation-invariant texture features from the outputs of two real Gabor filters in opposite symmetry to facilitate script recognition of six scripts: English, Chinese, Russian, Greek, Malayalam and Persian. This work was extended in Peake and Tan (1997) to include Korean language. Using Gabor energy features in a similar manner for script recognition purpose was also proposed by many researchers including Joshi et al. (2006). Next, the application of wavelet texture features in script recognition was introduced by Busch, Boles and Sridharan (2005). The authors concluded that using the wavelet log co-occurrence features could reduce the error rate of identifying between eight scripts: Latin, Chinese, Japanese, Greek, Cyrillic, Hebrew, Sanskrit and Farsi to 1.0% only. Another attempt to identify the script of text extracted from video and images was proposed by Gllavata and Freisleben (2005). They employed wavelet texture features to determine if the extracted text is Latin or ideographic (i.e. East Asian) achieving 87.20% accuracy rate. On top of that, Tho and Tang (2001) used the Modified Fractal Signature (MFS) and Modified Fractal Feature (MFF) in their script recognition method including English, Russian, French, German, Italian, Chinese, Japanese, Indian and Korean scripts. The overall average accuracy rate of 93.31% was obtained using a weighted Euclidean distance classifier.

HANDWRITTEN DOMAIN IN OSR APPLICATION

The application of local approaches in performing script recognition in handwritten domain was proposed by fewer researches as compared to that of machine-printed domain. Hochberg et al. (1999) adopted textual symbol analysis technique from their earlier work addressing machine-printed text. Hochberg et al. (1997) set statistical features that is extracted from document image via connected component labelling process, used to facilitate the script recognition process. The fisher linear discriminator is applied to differentiate between Arabic, Chinese, Cyrillic, Devanagari, Japanese, and Roman scripts with an overall accuracy of 88%.

Initially, the multi-script handwriting recognition motivated Sarkar et al. (2010) to develop an automatic script recognition for multi-OCR system. Thus, he presented script identification for two classes including Bangla and Devanagri-Roman script. They scored average accuracy rates 99.29% and 98.43% for Bangla and Devanagri-Roman script subsequently by using Multi-Layer Perceptron (MLP) as the classifier. With similar motivation, Pardeshi et al. (2014) achieved maximum accuracies up to 98% and 96% for bi-script and tri-scipt respectively after tested on Sarkar et al. (2010) dataset. Later, Obaidullah et al. (2015) claimed that designing multi-OCR system for handwritten was more challenging than machine-printed domain. Furthermore, the challenges became more complex in case of handwritten nested of machine printed. They proposed an automatic handwritten script identification for six Indian scripts by dividing into two parts which were 66% for training while the rest for testing. Their experimental result shows promising performance.

In conclusion, the script identification is a crucial step for automatic multi-lingual OCR system. The handwritten script document constitutes the real challenges, instead of machine printed, due to unconstraint nature of writing style. Rightfully, this paper centers upon the former where the most of multi-OCR's system are applied in multilingual environment. Human intervention is still required in existing multi-OCR system to channel the document into proper OCR. In this study, seven scripts are analysed for multi-OCR system automation perspective.

METHODOLOGY

Automatic multi-lingual script recognition (AMSR) is a pre-step of multi-optical character recognition (multi-OCR) system. The multi-OCR system usually appears in place where multilingual documents are gathered in mass collection such as international departments, tourism departments, postage companies, emigration departments and international airports. In this paper, an AMSR model is proposed based on both document image text thinning method for pre-processing phase, and texture analysis of binary document text image method for feature extraction phase. The basis on which the feature is extracted, is the Grey-Level Co-occurrence Matrix (GLCM) and Local Binary Pattern (LBP) after unifying the texture into block. In the recognition phase, the script recognition in multilingual documents text images entail the use of the multilayer neural network. The theoretical aspects of AMSR in multilingual document environment will be expounded in this section. Moreover, the whole system methodology is demonstrated and traced into automatic script recognition. It presents document image texture unification and methods used for each stage of AMSR.

MATERIAL

Scripts are a set of graphical representations used to express a particular writing system as well as subsets belonging to a particular writing system. The writing styles of more than one script family may then be adopted by one language, such as in the cases where the old Malay language (Jawi) adopts the Arabic script while the modern one adopts the Roman script. The seven major scripts used in this research are in handwritten style from Multilingual-HW benchmark dataset including Arabic, Devanagari, Hebrew, Thai, Greek, Cyrillic and Korean as shown in Figure 1. Over seven hundred binary image samples extracted represent considerable variations in writing styles.

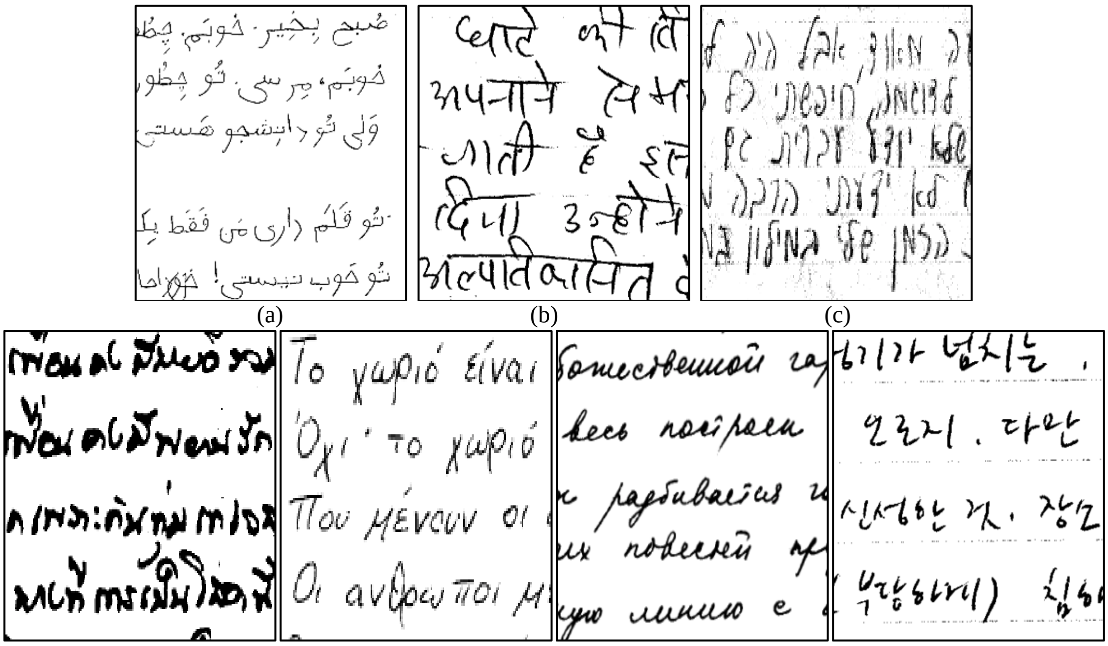
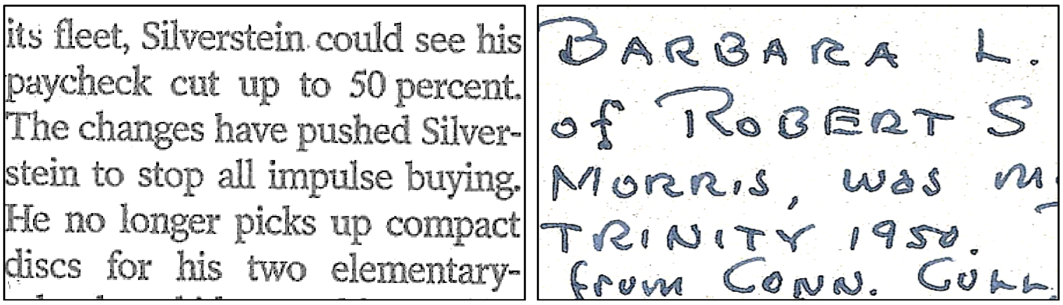

(a) (b) (c)
(d) (e) (f) (g)

FIGURE 1. Example of seven scripts used in this research; (a) Arabic, (b) Devanagari, (c) Hebrew, (d) Thai, (e) Greek, (f) Cyrillic, and (g) Korean

Scripts may be in the form of printed or written. The printed style is a machine-generated form depicted in Figure 2, which is formal, simple and used in printed matter such as newspapers and schoolbooks. Secondly, handwritten style depicted in Figure 2 (b), is usually the product of human-made writing in daily dealings. It is less formal and often used in ordinary documents.


its fleet, Silverstein could see his paycheck cut up to 50 percent. The changes have pushed Silverstein to stop all impulse buying. He no longer picks up compact discs for his two elementary-

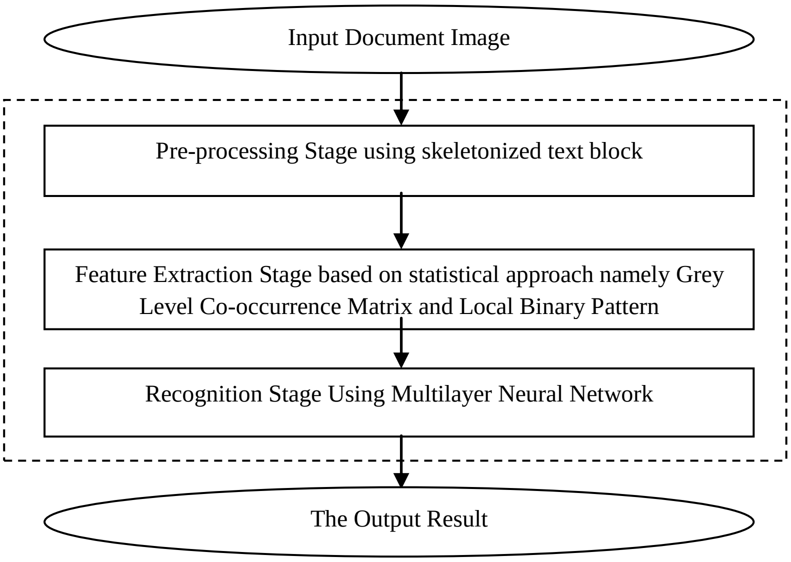
BARBARA L.
of ROBERT S
MORRIS, WAS M
TRINITY 1950.
from CONN. COLL

(a) (b)

FIGURE 2. (a). The printed Latin script style, and (b) the handwritten Latin style

THE PROPOSED AUTOMATIC OSR FRAMEWORK

The Automatic Multi-lingual Script Recognition (AMSR) framework stages comprise the pre-processing stage, feature extraction stage, and recognition stage. It is summarized and processed in Figure 3 and Algorithm 1 consecutively. The implementation of the framework for automatic multi-lingual script recognition system AMSR is aimed to cope with a multilingual environment document text image and tested on seven international scripts, which are Arabic, Devanagari, Hebrew, Thai, Greek, Cyrillic and Korean. Next, the description of each stage descript is detailed as follows.


Input Document Image
Pre-processing Stage using skeletonized text block
Feature Extraction Stage based on statistical approach namely Grey Level Co-occurrence Matrix and Local Binary Pattern
Recognition Stage Using Multilayer Neural Network
The Output Result

```
Algorithm 1; The proposed AMSR Framework
Input: binary document script image
Output: Script type
Begin
    1.  Create Skeletonized Text Block
            1.1  Skew correction
            1.2  text line detection
            1.3  line unification
            1.4  text block normalization
            1.5  thinning method
    2.  Extract Eighteen Features
            2.1  Apply the Grey-Level Co-occurrence Matrix (GLCM) and Local Binary Pattern (LBP) for
                 feature extraction methods
    3.  Script Type
            3.1  Feed the eighteen extracted features into several supervised classifiers including Bayer
                 Net, Decision Table, Random Forest, and Multilayer Neural Network.
            3.2  Multilayer neural network classifier gets the highest accuracy success rate (refer to Table
                 3).
End
```

FIGURE 3. The proposed AMSR framework

PRE-PROCESSING STAGE

Each input document comes with a variety of characteristics and properties, such as word spacing, text skewness, and text variant sizes. However, in pre-processing stage, a set of preparation steps are taken into account such as skew detection and correction, text normalization, and thinning method. The thinning method is performed to extract a text skeleton. Then, Hough transform method (Singh, Bhatia & Kaur, 2008) is applied to detect skewness angle for correction purpose. Text normalization eliminates word spaces, fills incomplete lines, and standardizes text size as well as the line number for text blocks generation. Next, the thinning method is performed on texture text blocks for skeleton extraction to be the input for the next stage. Finally, those thinned texture blocks or skeleton pixels will be represented via the feature extraction methods before exiting into recognition phase.

DOCUMENT IMAGE THINNING AND TEXTURE UNIFICATION

Myriads of document analysis and recognition applications require thinning process of document text images as a crucial step for representing the test pattern in one-pixel width skeleton, keeping all properties. The properties include the preservation connectivity, topology and sensitivity of text skeleton images. Image texture can be defined as a repeating pattern of pixels in a structured and consistent presentation. Text normalization for texture unification is still a fundamental problem in document text images for global feature extraction (Bataineh, Abdullah & Omar, 2011b). In a multilingual environment, a multi script considers that each script possesses its own properties in both writing style and physical formation. Evidently, Arabic script is cursive in nature whereby its characters are concatenated with each other and written from right to left. Whereas, the Latin characters are written separately and from left to right. These physical differences can result in significant distortion in the feature extraction. Based on the texture analysis techniques, hence, lies the importance of text normalization step to overcome these physical differences in synthesizing a uniformed texture black for global feature extraction. The skew correction, text line detection, line unification, text block unification and text block skeletonization are detailed sub-phases in pre-processing stage in which resultant a texture unification framework succinctly illustrated in Figure 4. Unification of multilingual scripts texture of skeleton undergoes the following steps:

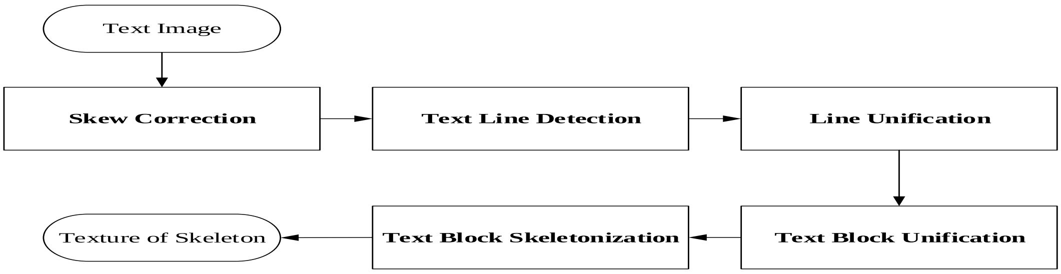


FIGURE 4. Texture unification processes in the pre-processing stage for the proposed AMSR application

FEATURE EXTRACTION STAGE

The aim of feature extraction based on statistical methods is to study the texture by computing the gray level values that represent occurrences (Rathore, 2014). Statistical methods are categorised into first-order, second and higher-order. We apply Grey-Level Co-occurrence Matrix (GLCM) which is a common algorithm created for texture features extraction in second-order statistic (Haralick & Shanmugam, 1973). GLCM manipulates and studies the relationship between two adjacent pixels in the original images in all directions (0°, 45°, 90°, 135°, 180°, 225°, 270°) and 315° using the following equation:

$$c_{\Delta x,\Delta y}(i,j) = \sum_{p=1,q=1}^{n,m} \begin{cases} 1, \text{if } I(p,q) = i \text{ and } I(p + \Delta x, q + \Delta y) = j \\ 0, \qquad \text{otherwise} \end{cases} \qquad (1)$$

where, *c* is a GLCM; *I* is an *n* × *m* image; (*Δx, Δy*) denotes the value of the pairs of pixels in particular direction, where the directions are represented by the 0°, 45°, 90°, 135°, 180°, 225°, 270° and 315°; (*i*, *j*) are the cell positions in the *c* matrix; and *p*, *q* are the values of the scoped pixels on *n*, *m*.

On top of that, Local Binary Pattern (LBP) operator is another second order algorithm developed to analyse texture of two-dimension surface via two factors; local spatial patterns and gray scale contrast (Ojala, Pietikäinen & Harwood, 1996). The original LBP operator forms the labels for the image pixels by thresholding the 3 × 3 neighbours of each pixel with the centre value and considering the result as a binary number. The LBP operator is extended to use neighbours of different sizes (Ojala, Pietikainen & Maenpaa, 2002). Using circular neighbours and bilinear interpolating values at non-integer pixel coordinates, any radius and number of pixels in the neighbours are allowed. The gray scale variance of the local neighbours can be used as the complementary contrast measure. Figure 5 shows the LBP calculation mechanism. Firstly, it examines the surrounding pixels of the centre value in the kernel matrix. The centre value shall be considered as the threshold value (for example pixel value > 6 is set as 1 or otherwise) for converting all surrounding pixels into binary format (Figure 5 (a)). Secondly, it concatenates a series of binary number starting from left of centre pixel into counter-clockwise direction as in Figure 5 (b). Finally, it converts the binary vector into decimal format to replace centre pixel × as in Figure 5 (c).

8	11	9
4	6	7
4	3	5

(a)

1	1	1
0	×	1
0	0	0

(b)

	Binary value	Decimal
×	00001111 (starting as in vector at (b))	15

(c)

FIGURE 5. An example of LBP calculation steps: (a) pixels values, (b) result based on threshold 6 (center), and (c) binary and decimal

In summary, statistical texture analysis method is regarded as one of the most effective among other texture analysis methods. Quevedo et al. (2013) claim that statistical texture approach is the most opted method in industrial fields geared towards classification purpose.

RECOGNITION STAGE USING MULTILAYER NEURAL NETWORK

Multilayer neural network comprises three stages, namely feed forward as input layer, the Back propagation association with error, and the adjustment of the weight, where the algorithm stages are demonstrated in Figure 6. The experiment settings used in all classifier are shown in Table 1.

```
Propagate the input forward through the network
        for every node in the layer
          1. Calculate the weight sum of the inputs to the node
          2. Add the threshold to the sum
          3. Calculate the activation for the node
        end
Propagate the errors backward through the network
      for every node in the output layer
        calculate the error signal
      end
      for all hidden layers
        for every node in the layer
          1. Calculate the node's signal error
          2. Update each node's weight in the network
        end
      end
Calculate Global Error
      Calculate the Error Function
        while ((maximum  number of iterations < specified) AND (Error Function is > specified))
```

FIGURE 6. The Backpropagation algorithm description (Sivanandam et al. 2006)

There in 18 nodes are set to input layer to represent the mean and standard deviation for direction angles for each primitive length of the skeleton and the primitive length of the skeleton for each particular direction angles respectively. The hidden layer contains 17 nodes, and 7 nodes in output layer representing script type as illustrated in Figure 7 whereas their parameters setting are mentioned in Table 1.

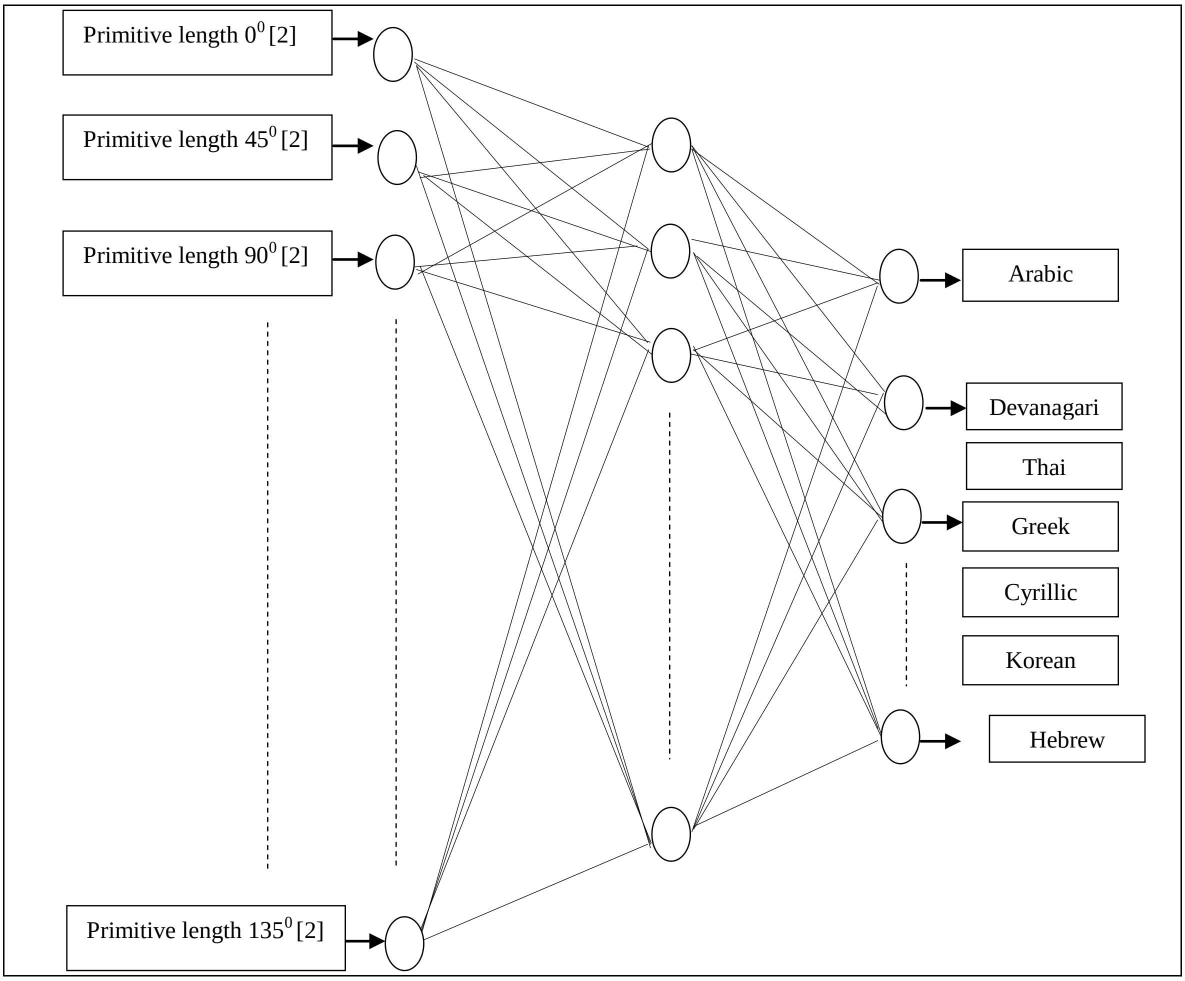

Primitive length 0^0 [2]
Primitive length 45^0 [2]
Primitive length 90^0 [2]
Primitive length 135^0 [2]
Arabic
Devanagari
Thai
Greek
Cyrillic
Korean
Hebrew

FIGURE 7. The structure of multilayer neural network which consists of input, hidden and output layers

TABLE 1. The detailed experimental settings of classifiers used in this research

Experiment setting for Multilayer Neural Network Classifier parameters	Percentage split	60%
	Number of training instances	308
	Number of Attributes	19
	Hidden layer	17
	Number of epochs	20
	Correctly Classified Instances	303
	Incorrectly Classified Instances	5
	Kappa statistic	0.981
	Mean absolute error	0.0136
	Root mean squared error	0.0587
	Momentum	0.3
	Relative absolute error	5.5751
	Root relative squared error	16.7738

EXPERIMENTAL RESULTS AND ANALYSIS

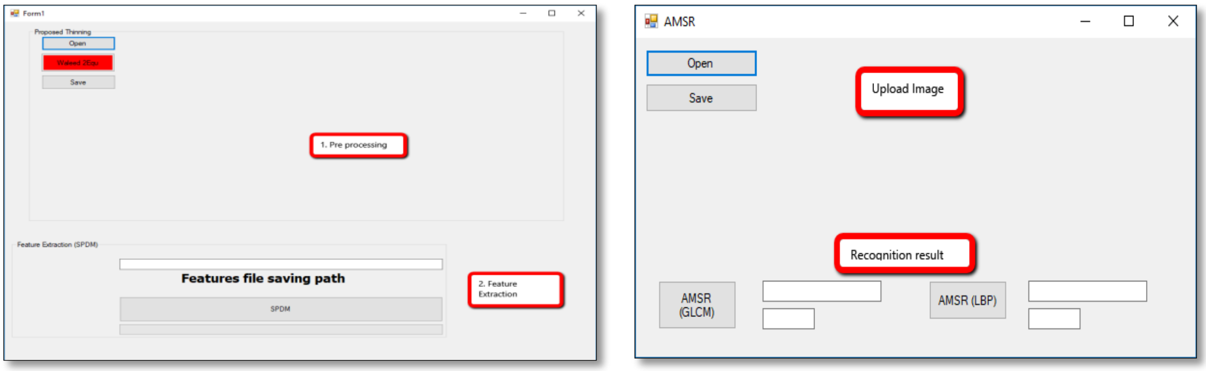
The general framework of proposed Automatic Multi-lingual Script Recognition (AMSR) was evaluated to measure its performance in **three** different types of experiments. **First**, the experiments were conducted based on the most powerful thinning methods on text image document as a step in pre-processing stage where the experiment was conducted on unified thinned text block of the multilingual-HW dataset. **Secondly**, the experiments laid the foundation for the analysis of the best texture statistical analysis method as a step in feature extraction stage. The performance of features were based on Grey-Level Co-occurrence Matrix (GLCM) and Local Binary Patterns (LBP). **Third,** the best methods were selected based on its respective performance in pre-processing stage and in feature extraction stage to be combined subsequently in classification stage. The overall AMSR was based on corresponding accuracy rate of the script type classification sample implemented using multilayer neural network.

EVALUATION OF THE AMSR BASED ON FEATURE EXTRACTION METHODS

The experiments were conducted on texture blocks of the Multilingual-HW dataset. This dataset was prepared by pre-processing to find the texture blocks of the international scripts. The final image results were in bitmap format with 512×512 pixel image size. The dataset consists of 700 image samples, covering 100 samples from each script of: Arabic, Devanagari, Hebrew, Thai, Greek, Cyrillic and Korean.

THE VISUAL EVALUATION

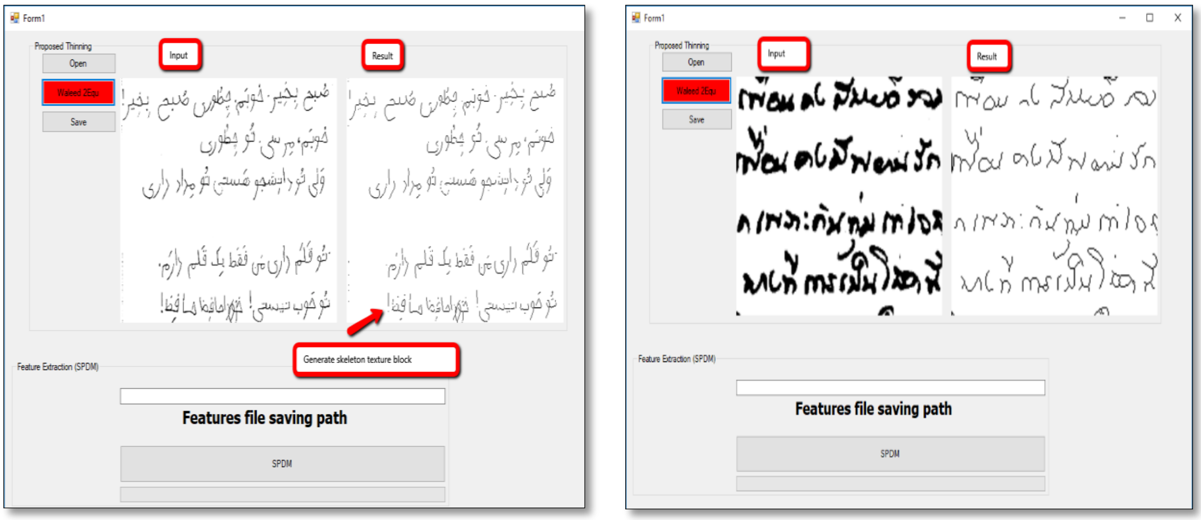
The performance of the proposed AMSR method was evaluated by comparing between different methods of feature extraction to select the best for the OSR. GLCM and LBP methods were conducted in these experiments because they were considered the most powerful and widely used as the feature extraction methods in many DIAR studies to analyse document surfaces and text typefaces, such as, script and language identification, writer identification and document characterisation and optical font recognition.



(a) (b)

FIGURE 8. The interface of: (a) Pre-processing and Feature extraction stage, (b) Recognition stage

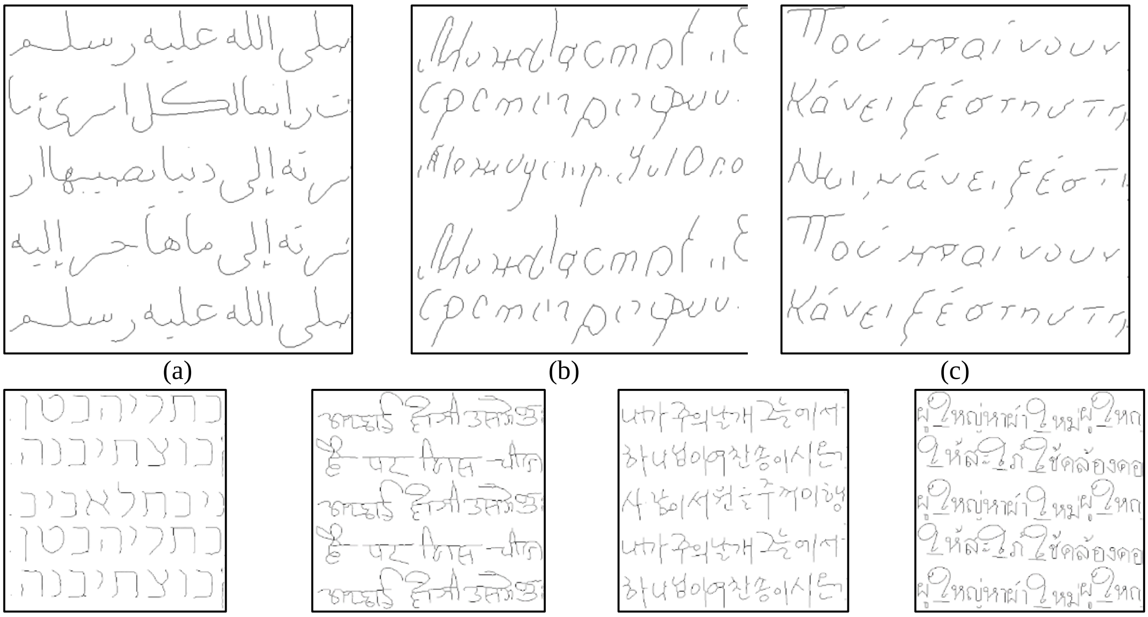
The pre-processing stage is to extract a one-pixel width skeleton for the text in all direction (tolerance to variant rotation angels). The output of pre-processing phase will be channelled into one of the most essential phase in pattern recognition application phases that is the feature extraction stage.


Form1
Proposed Thinning
Open
Waleed 2Equ
Save
Input
Result
Generate skeleton texture block
Feature Extraction (SPDM)
Features file saving path
SPDM

(a) (b)

FIGURE 9. The example of result for the document image text thinning method; (a) Arabic scripts, (b) Thai scripts

Figure 10 below shows the example overall visual results for each script class after completing the pre-processing stage. Skeletonization is able to extract single pixel information or skeleton of texture of each script class namely Arabic, Cyrillic, Greek, Hebrew, Devanagari, Korean, and Thai scripts.

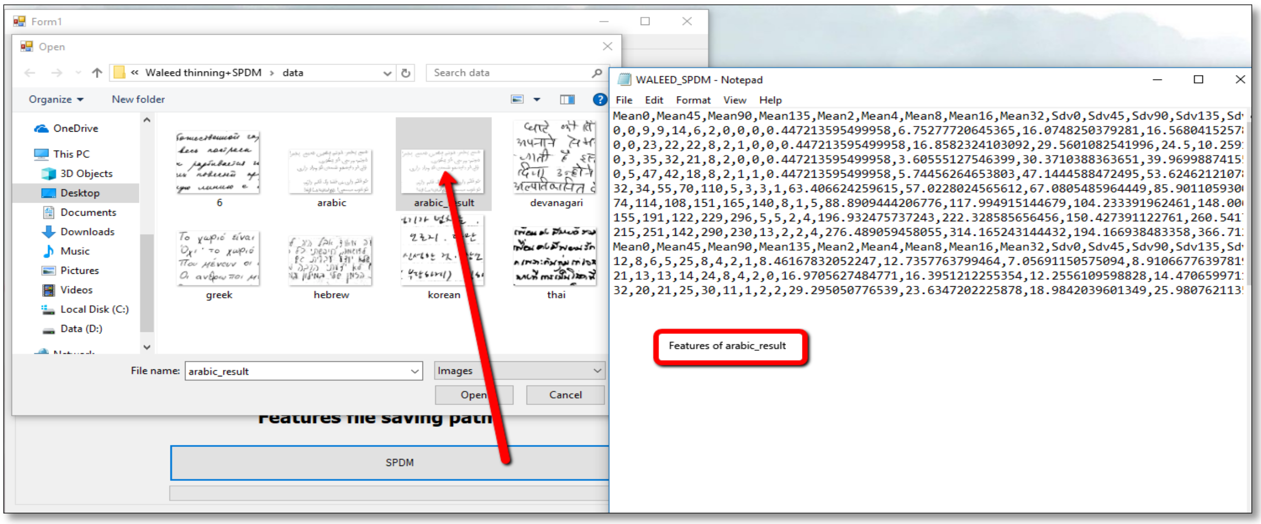


(a) (b) (c)

(d) (e) (f) (g)

FIGURE 10. The skeleton of the texture of: (a) Arabic, (b) Cyrillic, (c) Greek, (d) Hebrew, (e) Devanagari, (f) Korean, and (g) Thai scripts

Figure 11 shows the texture analysis of binary document text image method interface. The feature extraction embodies global approach which concerns with the extraction of global properties of an image. It is noteworthy that global feature extraction scope encompasses the overall or text bock level for image analyzing. This stage enables the generalisation into different application in document image analysis and recognition without major modification.



File name: arabic_result

```
Mean0,Mean45,Mean90,Mean135,Mean2,Mean4,Mean8,Mean16,Mean32,Sdv0,Sdv45,Sdv90,Sdv135,Sd
0,0,9,9,14,6,2,0,0,0,0.447213595499958,6.75277720645365,16.0748250379281,16.568041525
0,0,23,22,22,8,2,1,0,0,0.447213595499958,16.8582324103092,29.5601082541996,24.5,10.259
0,3,35,32,21,8,2,0,0,0.447213595499958,3.60555127546399,30.3710388363651,39.969988741
0,5,47,42,18,8,2,1,1,0.447213595499958,5.74456264653803,47.1444588472495,53.624621210
32,34,55,70,110,5,3,3,1,63.406624259615,57.0228024565612,67.0805485964449,85.90110593
74,114,108,151,165,140,8,1,5,88.8909444206776,117.994915144679,104.233391962461,148.00
155,191,122,229,296,5,5,2,4,196.932475737243,222.328585656456,150.427391122761,260.541
215,251,142,290,230,13,2,2,4,276.489059458055,314.165243144432,194.166938483358,366.71
Mean0,Mean45,Mean90,Mean135,Mean2,Mean4,Mean8,Mean16,Mean32,Sdv0,Sdv45,Sdv90,Sdv135,Sd
12,8,6,5,25,8,4,2,1,8.46167832052247,12.7357763799464,7.05691150575094,8.910667763978
21,13,13,14,24,8,4,2,0,16.9705627484771,16.3951212255354,12.2556109598828,14.47065997
32,20,21,25,30,11,1,2,2,29.295050776539,23.6347202225878,18.9842039601349,25.98076211
```

Features of arabic_result

FIGURE 11. The features developed from feature extraction stage

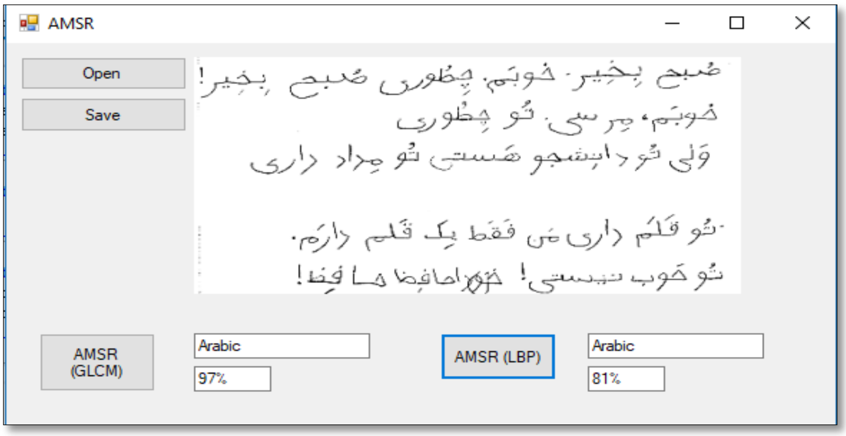
The objective of script type classification is to apply statistical features namely GLCM and LBP as the input to a set of classifiers such as bayes net, decision table, and multilayer neural network. The success rate of each classifier on 66% training dataset ratio on multilingual script is summarized as in Table 2. Moreover, the multilayer neural network manages to outperform other classifiers. This is due to the hidden layer in multilayer neural networks containing complex relationship between the input and output layers.

TABLE 2. Several classifiers are tested on Multilingual dataset on 66% training ratio

Classifier	**Bayes Net**	**Decision Table**	**Multilayer Neural Network**
Success Rate	85.12%	69.47%	98.87%.

The classification success rate achieved 98.87% through multilayer neural network classifier which utilizes 18 features to represent the input data, while the output is represented by seven scripts. To match the training experiments into multi neural network model, see Figure 11.

The script recognition in multilingual documents text images entails the use of the multilayer neural network. The result is based on GLCM and LBP features.



صُبح بخیر. خوبتم، چطوری صُبح بخیر!
خوبتم، مرسی. تُو چطوری
وَلی تُو دانشجو هستی تُو مداد داری

تُو قَلَم داری مَن فقط یک قلم دارَم.
تُو خوب نیستی! خداحافظ ما فقط!

Arabic
97%

Arabic
81%

FIGURE 12. The recognition stage interface

To present an objective evaluation of the quality of the extracted features, the proposed method GLCM and LBP methods were all applied on the same datasets, processes and a multilayer neural network classifier. The datasets were split into training and testing datasets. In this experiment, the training datasets were taken from a dataset of percentage 66%. As shown in Table 3, the accuracy rate of GLCM and LBP methods are 97.01% and 85.29% respectively with the multilayer neural network.

TABLE 3. The average of the classification results of the 66% training sets of the Multilingual-HW dataset using GLCM and LBP subsequently

	GLCM	**LBP**
Multilayer neural network	97.01	85.29

The confusion matrix results were extracted for scripts variation from experimental results of the proposed GLCM and LBP feature extraction methods. Based on the experiment, 66% percentage of training dataset and multilayer neural network classifier was selected for all tests. Based on confused matrix results of the GLCM for scripts identification shown in Table 4, the highest accuracy rate of 100% was achieved with the Hebrew script types, whereas the lowest accuracy was recorded at 94.8% for the Devanagari script sample.

TABLE 4. The confusion matrix of the GLCM with multilayer neural network classifier, Multilingual-HW dataset and 66% training dataset

	Arabic	**Devanagari**	**Hebrew**	**Thai**	**Greek**	**Cyrillic**	**Korean**
Arabic	96.97	0	0	0	0	0	3.03
Devanagari	5.13	94.87	0	0	0	0	0
Hebrew	0	0	100	0	0	0	0
Thai	0	0	0	97.73	0	0	2.27
Greek	0	0	0	0	96.97	3.03	0
Cyrillic	0	0	0	0	0	96.88	3.12
Korean	0	0	2.18	0	0	2.17	95.65

Based on confused matrix results of the LBP for scripts identification shown in Table 5, the highest accuracy rate of 91.20% was achieved with the Thai script types, whereas the lowest accuracy was at 79.6% for the Korean script sample.

TABLE 5. The confusion matrix of the LBP with multilayer neural network classifier, Multilingual-HW dataset and 66% training dataset

	Arabic	**Devanagari**	**Hebrew**	**Thai**	**Greek**	**Cyrillic**	**Korean**
Arabic	81.10	0	0	16.14	0	0	2.76
Devanagari	0	89.70	1.79	0	0	8.51	0
Hebrew	0	5.1	85.40	0	9.50	0	0
Thai	0	0	0	91.20	0	8.80	0
Greek	12.91	0	0	5.99	81.10	0	0
Cyrillic	0	0	3.24	0	0	88.90	7.86
Korean	0	19	0	6.72	0	13.68	*79.6*

For GLCM method, Arabic scripts were 5.13% recognized as Devaganari scripts whereas for LPB method, Cyrilic scripts were 13.68% recognized as Korean scripts. This is because the occurrences and relationship between the pixels with its neighbours among the scripts were almost approximate.

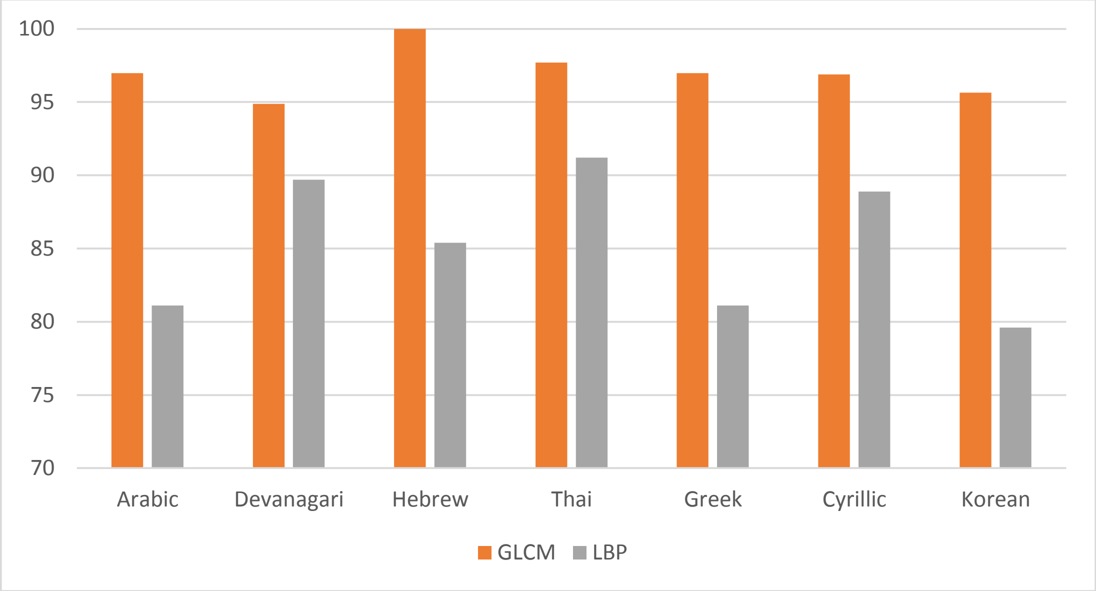


FIGURE 13. The correction rate for each class by GLCM and LBP features and multilayer neural network classifier, Multilingual-HW dataset with 66% training dataset

Notably, the GLCM method outperforms the LBP method for all scripts because of its ability to investigate the possible relationship of the primitive length in all directions. Both methods attempt to investigate the relationship between the examined pixels with their neighbour pixels. The static presentation values of binary format, which are zero and one, lead to the limitation in feature extraction possibilities such as occurrences and relationships of binary image pixels (Bataineh et al., 2011a; Bian, 2005).

SIGNIFICANCE OF THE WORK

Learning a new language through identifying multi-script recognition system is interesting because a person may be able to communicate across the world by firstly identifying their written or printed documents. Besides that, spoken and written language can also be associated to its ethnicity. In this study, we focus on seven interesting scripts namely Arabic, Devanagari, Hebrew, Thai, Greek, Cyillic and Korean because their scripts are less likely to be recognised in comparison to typical Latin scripts. Their written and printed text documents are looking similar to each other and very difficult to distinguish for a layman to grasp. This application also helps to reduce the language and communication gaps between races and countries. Besides that, this application is suitable to use when dealing with international languages such as at the international airport, postage company, international relations department, universal library, and email or web document language classification.

SUMMARY

In this study, an optical script recognition model is proposed. Seven international scripts were selected from multilingual dataset including Arabic, Latin, Hebrew, Devanagari, Greek, Cyrillic and Korean texts. Each step in the proposed framework is illustrated and presented including skew detection and correction, text line detection, text block unification, and text skeletonization based on the proposed thinning algorithm. After that, the proposed Grey-Level Co-occurrence Matrix (GLCM) and Local Binary Pattern (LBP) are used to extract features from the occurrences of the text primitive’s lengths and their orientation relationships. Both GLCM and LBP matrix use of statistical analysis include mean and standard deviation to extract 18 features.

In the recognition phase, the multilayer neural network classifier was chosen among variance classifiers, including Bayes Net, Decision Table, and Random Forest, based on its superior performance among a set of experiments. The eighteen features extracted using the proposed GLCM and LBP technique are represented by 18 nodes in the input layer of the multilayer neural network classifier yield 7 nodes represented in output layer as a script type. The proposed framework is able to deal with various number of scripts types without any major modification.

ACKNOWLEDGEMENT

The author would like to thank the Faculty of Information Science and Technology and Center for Research and Instrumentation Management of the Universiti Kebangsaan Malaysia for providing facilities and financial support. This work was funded by research grants PRGS 2016-1-ICT02/UKM/1/02/1 and AP 2017-005/2.

REFERENCES

A. Abidi, I. Siddiqi and K. Khurshid, (2011). "Towards Searchable Digital Urdu Libraries - A Word Spotting Based Retrieval Approach," *2011 International Conference on Document Analysis and Recognition*, Beijing, 1344-1348.

Ahmed, R., Al-Khatib, W. G. & Mahmoud, S. (2017). A survey on handwritten documents word spotting. *International Journal of Multimedia Information Retrieval. Vol. 6*(1), 31-47.

Bataineh, B., Abdullah, S. N. H. S. & Omar, K. (2011a). Generating an Arabic Calligraphy Text Blocks for Global Texture Analysis. *International Journal on Advanced Science, Engineering and Information Technology*. *Vol. 1*(2),50-155.

Bataineh, B., Abdullah, S. N. H. S. & Omar, K. (2011b). *A statistical global feature extraction method for optical font recognition.* Paper presented at the Asian Conference on Intelligent Information and Database Systems, 257-267.

Bataineh, B., Abdullah, S. N. H. S. & Omar, K. (2012). A novel statistical feature extraction method for textual images: Optical font recognition. *Expert Systems with Applications*. *Vol. 39*(5), 5470-5477.

Bataineh, B., Abdullah, S.N.H.S. & Omar, K. (2017) . Adaptive binarization method for degraded document images based on surface contrast variation. P*attern Analysis Applications*. *Vol. 20*(3), 639-652.

Bian, N. (2005). *Evaluation of Texture Features for Analysis of Ovarian Follicular Development.* Master thesis, University of Saskatchewan, Saskatoon.

Boufenar, C., Kerboua, A. & Batouche, M. (2018). Investigation on deep learning for off-line handwritten Arabic character recognition. *Cognitive Systems Research. Vol. 50*(180-195).

Busch, A., Boles, W. W. & Sridharan, S. (2005). Texture for script identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 27*(11), 1720-1732.

Chen, H., Tsai, S. S., Schroth, G., Chen, D. M., Grzeszczuk, R. & Girod, B. (2011). *Robust text detection in natural images with edge-enhanced maximally stable extremal regions*. 2011 18th IEEE International Conference on Image Processing, Brussels, 2609-2612.

Ghosh, D., Dube, T. & Shivaprasad, A. (2010). Script recognition—a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 32*(12), 2142-2161.

Gllavata, J. & Freisleben, B. (2005). *Script recognition in images with complex backgrounds.* Paper presented at the Signal Processing and Information Technology, 2005. Proceedings of the Fifth IEEE International Symposium on., 589-594.

Haralick, R. M. & Shanmugam, K. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics. Vol. 6*, 610-621.

Hochberg, J., Bowers, K., Cannon, M. & Kelly, P. (1999). Script and language identification for handwritten document images. *International Journal on Document Analysis and Recognition. Vol. 2*(2), 45-52.

Hochberg, J., Kelly, P., Thomas, T. & Kerns, L. (1997). Automatic Script Identification From Document Images Using Cluster-Based Templates. *IEEE Trans. Pattern Anal. Mach. Intell. Vol. 19*(2), 176-181. doi: 10.1109/34.574802

Jain, A. K. & Zhong, Y. (1996). Page segmentation using texture analysis. *Pattern Recognition. Vol. 29*(5), 743-770.

Jiang, X. (2009). "*Feature extraction for image recognition and computer vision*". Paper presented at the 2009 2nd IEEE International Conference on Computer Science and Information Technology,1-15. 8-11 Aug.

Joshi, G. D., Garg, S. & Sivaswamy, J. (2006). *Script identification from Indian documents.* Paper presented at the Document Analysis Systems. 255-267.

Kamble, P. M. & Hegadi, R. S. (2015). Handwritten Marathi character recognition using R-HOG Feature. *Procedia Computer Science. Vol. 45*, 266-274.

Kasturi, R., O'gorman, L. & Govindaraju, V. (2002). Document image analysis: A primer. *Sadhana. Vol. 27*(1), 3-22.

Khaleefah, S. H. & Nasrudin, M. F. (2016). Identification of printing paper based on texture using gabor filters and local binary patterns. *Journal of Theoretical and Applied Information Technology. Vol. 86*(2), 279-289.

Li, J., Fan, Z.-G., Wu, Y. & Le, N. (2009). *Document image retrieval with local feature sequences.* Paper presented at the Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on. 346-350.

Lutf, M., You, X., Cheung, Y.-m. & Chen, C. P. (2014). Arabic font recognition based on diacritics features. *Pattern Recognition. Vol. 47*(2), 672-684.

Marinai, S. (2008). Introduction to document analysis and recognition. *Machine Learning in Document Analysis and Recognition.* 1-20.

Obaidullah, S. M., Das, N., Halder, C. & Roy, K. (2015). Indic script identification from handwritten document images — An unconstrained block-level approach," 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), Kolkata, 2015, 213-218.

Ojala, T., Pietikäinen, M. & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition. Vol. 29*(1), 51-59.

Ojala, T., Pietikainen, M. & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 24*(7), 971-987.

Pardeshi, R., Chaudhuri, B., Hangarge, M. & Santosh, K. (2014). *Automatic handwritten Indian scripts identification.* Paper presented at the Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. 375-380.

Peake, G. & Tan, T. (1997). *Script and language identification from document images.* Paper presented at the Document Image Analysis, 1997.(DIA'97) Proceedings., Workshop on.,10-17.

Peete, B. P. & A. G. Ramakrishnan. (2008). Word Level Multi-Script Identification. *Pattern Recognition Letters. Vol. 29,* 1218-1229.

Quevedo, R., Valencia, E., Bastías, J. M. & Cárdenas, S. (2013). *Description of the enzymatic browning in avocado slice using GLCM image texture.* Paper presented at the Pacific-Rim Symposium on Image and Video Technology. 93-101.

Radwan, M. A., Khalil, M. I. & Abbas, H. M. (2017). Neural networks pipeline for offline machine printed Arabic OCR. *Neural Processing Letters*. 1-19.

Rao, G. S., Imanuddin, M. & Harikumar, B. (2014). Script Identification of Telugu, English and Hindi Document Image. *Int. J. Adv. Eng. Global Technol. 2*(2), 443-452.

Rathore, M. S. (2014). *Statistical analysis of Synthetic Aperture Radar (SAR) image speckle.* Retrieved from Biju Patnaik Central Library National Insitute of Technology Rourkela, Odisha-769008, 5946.

Saabni, R., Asi, A. & El-Sana, J. (2014). Text line extraction for historical document images. *Pattern Recognition Letters. Vol. 35*, 23-33.

Sarkar, R., Das, N., Basu, S., Kundu, M., Nasipuri, M. & Basu, D. K. (2010). Word level script identification from Bangla and Devanagri handwritten texts mixed with Roman script. *arXiv preprint arXiv:1002.4007*.

Singh, C., Bhatia, N. & Kaur, A. (2008). Hough transform based fast skew detection and accurate skew correction methods. *Pattern Recognition. Vol. 41*(12), 3528-3546.

Singh, R., Yadav, C., Verma, P. & Yadav, V. (2010). Optical character recognition (OCR) for printed devnagari script using artificial neural network. *International Journal of Computer Science & Communication. Vol. 1*(1), 91-95.

Sulaiman, A., Omar, K. & Nasrudin, M. F. (2017). A database for degraded Arabic historical manuscripts. *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*. 1–6.

Ubul, K., Tursun, G., Aysa, A., Impedovo, D., Pirlo, G. & Yibulayin, T. (2017). Script Identification of Multi-Script Documents: A Survey. *IEEE Access. Vol.* 5, 6546–6559.

Tan, T. (1998). Rotation invariant texture features and their use in automatic script identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 20*(7), 751-756.

Tan, X. & Triggs, B. (2010). Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing. Vol. 19*(6), 1635-1650.

Tensmeyer, C. & Martinez, T. (2017). Document Image Binarization with Fully Convolutional Neural Networks. *arXiv preprint arXiv:1708.03276.*

Tho, Y. & Tang, Y. Y. (2001). *Discrimination of oriental and Euramerican scripts using fractal feature.* Paper presented at the Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on. 1115-1119.

Vinod, H. & Niranjan, S. (2018). *Multi-level Skew Correction Approach for Hand Written Kannada Documents.* Paper presented at the International Conference on Information Theoretic Security. 376-386.

Zavvar, M., Garavand, S., Nehi, M. R., Yanpi, A., Rezaei, M. & Zavvar, M. H. (2016). Measuring Reliability of Aspect-Oriented Software Using a Combination of Artificial Neural Network and Imperialist. *Asia-Pacific Journal of Information Technology and Multimedia. Vol.* 5(2), 75-84.

ABOUT THE AUTHORS

Waleed Abdel Karim Abu-Ain is a PhD student at the Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia. He is from King Abdulaziz University, Saudi Arabia.

Siti Norul Huda Sheikh Abdullah is a senior lecturer at Center for Cyber Security, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia.

Khairuddin Omar is a senior lecturer at Center for Artificial Intelligence Technology, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia.

Siti Zaharah Abd. Rahman is a senior lecturer at Center for Cyber Security, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia.