International Journal of Computer Applications Technology and Research
Volume 4– Issue 2, 115 - 118, 2015, ISSN:- 2319–8656

Haralick Texture Features based Syriac(Assyrian) and English or Arabic documents Classification

Basima Z.Yacob
Department of computer science
Faculty of science
University of Duhok
Duhok, Iraq

Abstract: Script identification is very essential before running an individual OCR system. Automatic language script identification from document images facilitates many important applications such as sorting, transcription of multilingual documents and indexing of large collection of such images, or as a precursor to optical character recognition (OCR), in this paper the characterized are between Syriac and English documents or between Syriac and Arabic documents were the characterized is achieved by extracting Haralick texture Features. it is investigated a texture as a tool for determining the script of document image ,based on the observation that text has a distinct visual texture. Further, K nearest neighbour algorithm is used to classify 300 text blocks into one of the two scripts: Syriac, and English , or Syriac and Arabic based on Haralick texture Features . The script was inserted to the System with different rotation angles between 0º and 135º and the results of recognition were good.

Keywords: Syriac script; Haralick Texture Features; OCR; English script; Arabic script; knn algorithm.

1. INTRODUCTION

Script and language identification are a key part of automatic processing of document images in an international environment. A document's script must be recognized in order to choose an appropriate optical character recognition (OCR) algorithm. For scripts used by more than one language, discriminating the language of a document prior to OCR is also helpful, and language identification is crucial for further processing steps such as routing, indexing, or translation.

One of the important tasks in machine learning is the electronic reading of documents. All documents can be converted to electronic form using a high performance Optical Character Recognizer (OCR). Recognition of bilingual documents can be approached by the recognition via script identification.

This paper considers the discrimination between the Syriac and English scripts and between the Syriac and Arabic scripts according to an analysis of text block.

The Syriac (Assyrian) language is one of the Semitic languages that is being spoken in Iraq, Syria, Turkey and Iran by Assyrians. It's an ancient language, one of the rarest and oldest in the world.

Syriac is an ancient Iraqi language, and it is culturally used by human beings in Iraq. It has many religious scripts as well as scientific and literary books which have been completed and achieved throughout the long history and efficient civilization for this language, and conveying this important thought for communication between the present and past generations.

Over the past decades, many different researches and papers have been concerned to discriminate between the two or more difference languages for example Arabic and English or between Indian and English documents and ect. , but no research has been achieved towards the discriminating between Syriac and other languages.

This paper presents a scheme for identification between Syric and English scripts or between Syriac and Arabic script based on Haralick Texture Features.

Two scripts were classified by the classification algorithm, these scripts are Syriac and Roman (English) or Syriac and Arabic. Classification accuracy depends on the rotation angle of the script.

2. RELATED WORK

Santanu Choudhuri, et al. [1] has proposed a method for identification of Indian languages by combining Gabor filter based technique and direction distance histogram classifier considering Hindi, English, Malayalam, Bengali, Telugu and Urdu. Dhanya et al. [2] have used Linear Support Vector Machine (LSVM), K-Nearest Neighbour (K-NN) and Neural Network (NN) classifiers on Gabor-based and zoning features to classify Tamil and English scripts. Wood et al. [3] have proposed projection profile method to determine Roman, Russian, Arabic, Korean and Chinese characters.

Later, script recognizer[4] has been extended to four scripts/ languages (Kannada, Hindi, English and Urdu) with different font sizes and styles by relaxing their constraints over different font sizes[5].

Horizontal projection was attempted [6] to separate two languages English and Arabic at text line level. Here, the horizontal projection profiles of Arabic text have a single peak corresponding to the baseline of the Arabic writing, where characters are connected together. In contrast, projections of English text have two major peaks corresponding to x-line and baseline. The projections of Arabic text lines are smooth while the projections of English text line have sharp jumps Multichannel Gabor filtering designed with four frequencies and four orientations was also applied over the bilingual document images[7] Arabic-English, Chinese-English, Hindi-English and Korean-English bilingual dictionaries to identify the script at word level.

Using the combination of shape, statistical and Water Reservoirs, an automatic line-wise script identification scheme from printed documents containing five most popular scripts in the world, namely Roman, Chinese, Arabic, Devnagari and Bangla has been introduced[8].

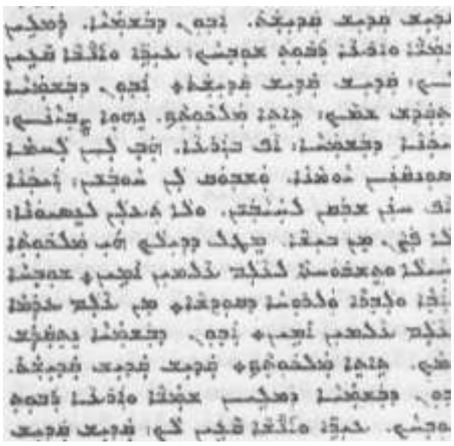
3. PROPOSED APPROACH

The paper primarily aims for block level classification; blocks of text are first extracted from the scanned document. For block of text extracted, Haralick Texture features are computed. These features are integrated to form database of vectors which are then used for Syriac and English or Arabic text separation via k-NN classifier. For better understanding, Figure.1 shows a schematic work-flow of the system

Figure. 1 A screen-shot of an overall work-flow of the system

3.1 Preprocessing

The preliminary task is to do pre-processing. Pre-processing techniques are application dependent. In this paper initially, 600 x600 text blocks are segmented manually from the document images of Syriac, English and Arabic and created 300 text blocks. Out of these 300 images Syriac, English, and Arabic are 100 each. A sample images text blocks of Syriac , English and Arabic are shown in figure 2.



(a)

(b)

(c)

Figure. 2 Examples of document images used for training and testing.

(a) Syriac, (b) English, and (c) Arabic.

3.2 Haralick TEXTURE Features EXTRACTION

From each block of normalized text, the Haralick texture features are evaluated for the purpose of script identification.
Haralick Texture features are first reported in [9] for image classification. For better understanding, texture can also be defined as:it is property which contains important information about structural arrangement of surfaces and their relationship with surrounding environment. In this paper, the Haralick Texture of each test image is extracted as attributes to build a database which is used at classification stage. These set of statistical texture features collectively used to generate a feature vector.
Haralick features are used for analyzing the texture of an image on the other hand; Haralick features offer 13 different elements that define the textural structure of a image. Haralick features can be defined as follows [9].
Contrast, Homogeneity, Dissimilarity, Energy and Entropy, as Angular second moment: Energy

$$f_1 = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} \{p(i.j)\}^2$$

Contrast:

$$f_2 = \sum_{n=0}^{Ng-1} n^2 \left(\sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p(i,j) \right)$$
$$\text{when } \lfloor i-j \rfloor = n$$

Correlation:

$$f_3 = \frac{\sum_{i=1}^{Ng} \sum_{J=1}^{Ng} (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \, \sigma_y}$$

Sum of squares: Variance

$$f_4 = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} (i-\mu)^2 \ p(i,j)$$

Inverse Difference Moment homogeneity
Homogeneity (HOM) (also called the "Inverse Difference Moment")

$$f_5 = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} \frac{1}{1+(i-j)^2} p(i,j)$$

Sum Average

$$f_6 = \sum_{i=2}^{2Ng} i p_{x+y}(i)$$

Sum Variance

$$f_7 = \sum_{i=2}^{2Ng} (i - f_8)^2 \ p_{x+y}(i)$$

Sum Entropy

$$f_8 = -\sum_{i=2}^{2Ng} p_x + y^{(i)} \log\{p_x + y^{(i)}\}$$

Entropy

$$f_9 = -\sum_{i=1}^{Ng}\sum_{j=1}^{Ng} p(i,j) \ \log(p(i,j))$$

Difference Variance

$$f_{10} = E[p_x - y^2] - E[p_x - y]^2$$

Difference Entropy

$$f_{11} = -\sum_{i=0}^{Ng-1} p_{x-y^{(i)}} \log\{p_{x-y^{(i)}}\}$$

Information Measures of Correlation

$$f_{12} = \frac{HXY - HXY1}{\max\{HX, HY\}}$$

$$f_{13} = (1 - \exp[-2.0(HXY2 - HXY)])^{1/2}$$

3.3 Classification

The traditional and simplest classification algorithm is k-nearest neighbour algorithm (k-NN). It is a method of classifying the instances based on the nearest training examples in the feature space. It classifies an object based on a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours. The training set includes the data for classification for each specific.

For every new input, the Haralick textural features are obtained. A sample of Haralick textural features of Syriac, English and Arabic scripts of figure 2 are represented in Table1.

The following are the steps of the algorithm

1. Given an input image X with different rotation angles between 0º and 135º, determine its distance measure based on the computation of textural features.
2. Determine the k (k=3) nearest neighbor in the training set which comprises of the Haralick features.
3. Assign the image X to the closest match.

Table1: The sample Haralick Texture Features of Syriaic, English and Arabic Scripts

script / Features	Syriac	English	Arabic
F1	0.6343	0.3782	0.5194
F2	0.2930	0.7979	0.2831
F3	175.4046	246.0818	221.0589
F4	14.2433	14.4900	16.3956
F5	0.9252	0.7981	0.8971
F6	7.4427	7.4751	8.0417
F7	45.6203	40.2149	50.4000
F8	0.8193	1.2732	1.0232
F9	1.0219	1.7562	1.2327
F10	0.0935	0.0547	0.0801
F11	0.4700	0.8992	0.5703
F12	-0.4320	-0.1738	-0.2773
F13	0.6561	0.5331	0.5725

4. DISCUSSION

Experimentations are carried out with KNN classifier. To evaluate the a sample image of size 600x600 pixels is selected manually from each document image and created 300 text block images. Out of these 300 images Syriac, English, and Arabic are 100 each. The accuracy of the classification achieved for script identification is shown in Tables 2 and 3.

The achieved results of the classification depend on the rotation angle of script.

Table 2. Text block Syriac-English scripts identification results

Type of Documents Syriac –English	No. of documents	Classified correctly	% correct classification
Syriac – with rotation 0˚	100	100	100%
Syriac – with rotation 45˚	100	100	100%
Syriac – with rotation 90˚	100	100	100%
Syriac – with rotation 135˚	100	100	100%
English – with rotation 0˚	100	75	75%
English – with rotation 45˚	100	0	0%
English – with rotation 90˚	100	75	75%
English – with rotation 135˚	100	0	0%

5. ACKNOWLEDGMENTS

My thanks to my late husband **Hormuz Bobo** who had passed away before I publish this paper, and he had contributed towards this paper and supported me and provided the Assyrian (Syriac) scripts but I thank him mostly because he always encouraged me to accomplish such projects about the Assyrian language which are all dedicated to him.

Table 3. Text block Syriac-Arabic scripts identification results

Type of Documents	**No. of documents**	**Classified correctly**	**% correct classification**
Syriac –Arabic			
Syriac – with rotation 0˚	100	100	100%
Syriac – with rotation 45˚	100	100	100%
Syriac – with rotation 90˚	100	100	100%
Syriac – with rotation 135˚	100	100	100%
Arabic– with rotation 0˚	100	100	100%
Arabic– with rotation 45˚	100	0	0%
Arabic– with rotation 90˚	100	100	100%
Arabic– with rotation 135˚	100	0	0%

6. REFERENCES

[1] Santanu C, Gaurav H., Shekar M.i, and Shet R.B., 2000, Identification of scripts of Indian languages by Combining trainable classifiers, Proc. of ICVGIP, India.

[2] Dhanya D., Ramakrishnan A.G. and Pati P.B., 2002, Wavelet Based Co-occurrence Histogram Features for Texture Classification with an Application to Script Identification in a Document Image, Pattern Recognition Letters 29, 2008, pp 1182-1189.

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

[4] Basavaraj P. and Subbareddy N. . Neural network based system for script identification in Indian documents, Sadhana Vol. 27, part-i1, pp 83-97, 2002.

[5] Dhandra.B.V, Nagabhushan. P, Mallikarjun H. , Ravindra H., Malemath. V.S, 2006. Script Identification Based On or phological Reconstruction In Document Images, The 18th International Conference on Pattern Recognition (ICPR’06).

[6] Elgammmal.A.M and Ismail.M.A, 2001. Techniques For Language Identification for Hybrid Arabic-English Document Images, Proc. Sixth Int’l Conf. Document Analysis and Recognition, pp. 1100-1104.

[7] Huanfeng M. and David D., 2003. Gabor Filter Based Multi-Class Classifier for Scanned Document Images, Proceedings of the Seventh International Conference on Document Image Analysis and Recognition (ICDAR’03).

[8] Pal U. and Chaudhuri.B.B, 2001., Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line, Proc. 6th Intl. Conf: Document Analysis and Recognition (ICDAR'OI), pages 790-794.

[9] R. M. Haralick, K. Shanmugam, I. Dinstein, 1973. Textural features for image classification, IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC 3, No.6, November, pp. 610-621.