International Journal of Engineering Works
Kambohwell Publisher Enterprises
ISSN: 2409-2770 Vol. 5, Issue 4, PP. 56-60, April 2018
www.kwpublisher.com

Classification Performance of Linear Binary Pattern and Histogram Oriented Features for Arabic Characters Images: A Review

Sungin Behram Khan, Dr. Gulzar Ahmad, Faheem Ali, Farooq Faisal, Irfan Ahmed, Salman Elahi

***Abstract*—**There are millions of texts store in both off line and online forms. To utilize these documents properly, there is need of organizing these documents systematically and lots of applications are available for this purpose. Text classification is an important area of image processing deal with how the document belongs to its suitable class or category. Like other languages, Arabic language is also very rich and complex inflectional language which makes Arabic language very complex for ordinary analysis. In this review paper, we focus on the published research, especially in the field of Arabic text classification. Regard these all, three different types of feature extraction techniques are also implemented to extract features from different images of Arabic characters and presents a performance results of these techniques. From the result, it can be concluded that the combination of Linear binary pattern descriptor and Legendre moment, based moments features outperform and increase the accuracy of the LBP classifiers from 91.99 % to 93.12%.

***Keywords*—** Text classification, Local Binary Pattern descriptor, Histogram of Gradient Feature descriptor, Legendre Moment, Classification.

I. INTRODUCTION

Text classification is a technique to extract useful information from the large amount of textual data. In the last few years, there is a dynamically growing of textual data in various fields of daily life, which made the text classification one of the most important research issues. Text classification is an active research area in which different labels are assigned to text documents with different categories from a predefined set known in advance [1]. There are two different terms in text classification, the first term is text categorization which deals with sorting documents by contents, while the term text classification is used to classify the characters [2]. There are two methods for the classification of text data: rule base and machine learning [3]. In the rule based, a system automatically classifies a textual data with the help of built in knowledge engineer and a domain expert while in the machine learning approaches, a system is fed to a set of data for training a machine and classes are defined for each different character. And new data are classified according to their distant properties [4]. Like other languages, the Arabic language has also very rich morphology and have complex shapes of their character [5]. Due to complex writing styles of Arabic characters, there is a decreased research in the area of Arabic text classification [6]. There are 28 letters in Arabic language [7]. The quality of the data set of data may very affect the classification performance of a machine; the redundant and irrelevant features of data may also reduce the classification performances of any algorithms [3]. For English text, there are seven different standards of text available free for research purposes, but for Arabic text classification unfortunately there is no specific standard data set [8]. Researches in the field of text classification collected their data from the online web sites [9]. The rest of the paper is organized as follows. Section 2 presents a brief summary of the previous work in the field of Arabic text classification follow by section 3 in which different feature extraction techniques are presented. Section 4 presents the classification and types of classifiers used in this research. In section 5 results of different features extraction techniques and classifiers are presented and 7 presents the conclusion.

II. RELATED WORK

Handwritten character recognition is one of the most challenging research directions (area) in artificial intelligence and pattern recognition [10, 11]. Character recognition helps you to transform different historical books, inscription, newspaper and unrestricted document formats to an intelligible format. It is almost impossible to manually annotate all of these handwritten documents. Therefore, there is a need of automatic system for labeling the words in the handwritten documents. Furthermore, the character recognition systems are related to research areas such as writer identification and verification [10]. Arabic handwriting recognition was first performed by [12] which is based on Fourier transformations. In [12], features are extracted from both isolated characters and cursive characters. The author classifies 175 samples of hand printed letters correctly. [13] introduced a method based on the

Sungin Behram Khan: Department of Electrical Engineering, University of Engineering and Technology Peshawar, Pakistan.

Dr. Gulzar Ahmad:Department of Electrical Engineering, University of Engineering and Technology Peshawar, Pakistan.

Faheem Ali: Department of Electrical Engineering, University of Engineering and Technology Peshawar, Pakistan.

Farooq Faisal: Department of IBMS Agriculture University Peshawar.

Irfan Ahmed: Department of Electrical Engineering, University of Engineering and Technology Peshawar, Pakistan.

Salman Elahi: Department of Electrical Engineering, University of Engineering and Technology Peshawar, Pakistan.

statistical approach for the recognition of isolated Arabic characters. Character recognition system is mainly divided into different types according to their techniques of data/image acquisition methods. Arabic recognition system is mainly divided into two types, i.e. Online and offline line systems. Some of these online systems [5] [6] get character data in real time. [19] develop a recognition method which segment the cursive words into different characters. Some recognition systems recognize the text words without segmenting it into words, characters or primitives [1, 7]. Images of separate characters are used in our experiments.

III. FEATURE EXTRACTION

Feature extraction is the process of extracting information from input image of the character or word and used that information for the recognition of the character. These extracted information is called feature vectors. It is most the important step in developing recognition system. Different types of feature extraction techniques are used to extract feature vector from images of characters. Some of these techniques are pixel-based method [14]. In this research paper, different feature extraction techniques, namely the Linear Binary Pattern descriptor (LPP),

Histogram of Gradient descriptor (HOG) and combined features of Legendre moment, based and local binary pattern are used to classify different Arabic characters. These descriptors are further explained in the coming next sections.

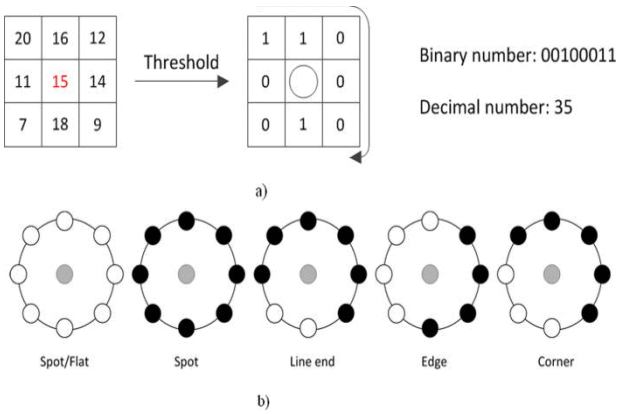

Figure 1. (a) The LBP operator. (b) Examples of texture primitives which can be detected by LBP, where white circles represent ones and black circles zeros. For instance, the rightmost pattern detects corners in an image

A. Local Binary Pattern Descriptor

The Local Binary Pattern descriptor introduced by [15], is a regional descriptor-based approach for texture description. It was later introduced into the face recognition area by [16]. The LBP generation approach in [15] is described as follows. A Local Binary Pattern label is a binary number which is created for each pixel of images where each bit is assigned a value based on its difference from one of the pixels at a given radius. Figure 1 (a) shows a simple version that labels each pixel by a number derived from its 3x3 neighborhood. Neighboring pixels are assigned a binary value of 1 if larger than the center pixel and 0 otherwise. Figure 1 (b) shows how different bits can be clustered to representative's spots, edges, lines, corner and edges. Locations of any pixel having not integer coordinates, these coordinates are calculated by using bilinear interpolation of each neighboring pixels. The feature vectors are from each section is obtained from the histograms of the LBP-operated image sections.

The feature vector assignment algorithm steps consist of different steps.

• Labeling each pixel using the LBP operator

• Dividing the image into m × m small equal sized rectangular regions R0R1,R(m□1)

• Obtaining the histogram from each computed region

• Combining all calculated histograms into one vector

Let f (x, y) be the LBP label of pixel of images at coordinates (x, y). Then the histogram for this region j can be computed using equation below.

$$H_{i,j} = \sum_{x,y} I[f(x,y) = i] I(x,y) \in R_j \quad (1)$$

where $i = 0, 1, ..., n - 1$, $j = 0, 1, ..., m - 1$ and

$$I\{A\} = \begin{cases} 1, & A \text{ is true} \\ 0, & A \text{ is false} \end{cases} \quad (2)$$

Histograms computed by equation 2 are estimates of the density function of the regional data which contains information about the patterns there in each pixel. i is a histogram domain that corresponds to one of the n = 2P binary numbers generated by LBPPR in the neighborhood of P pixel. The resultant feature vectors are very high even for typical value of P. For this some clustering techniques can be used to reduce the feature vector size.

B. Histogram Of Gradient Features Descriptor

Dalal et al .[17] introduced Histogram Oriented Gradient for the detection of the human body in an image. It has become very successful in diverse domains such as face detection [18], pedestrian detection [19] and road vehicle detection [20]. HOG is defined as the distribution of the local intensity gradients over a small connected region called "cells". Feature vector is computed for each cell using gradient detectors. Each pixel is then convolved by simple convolution kernel as:

$$Gx = I(x + 1, y) - I(x - 1, y) Gy = I(x, y + 1) - I(x, y - 1) \quad (3)$$

In equation 3, I (x, y) is the pixel intensity at location x, y. Gx are the horizontal component and Gy are the vertical components of the gradients. Feature are represented by the combination of the histograms from each block. The size of the feature vector in HOG descriptor depends on the selected numbers of bins and blocks. The performance of the HOG descriptor is mostly depending on the number of selected bins and blocks [19]. The feature descriptors can be normalized using Level 2 block normalization [18] as follows:

$$V_k' = \frac{V_k}{\sqrt{\|V\|^2 + \varepsilon}} \quad (4)$$

ISSN: 2409-2770

- In equation 4, Vk represent the combined histogram from all block regions, ε represent small value close to zero and V′
- k is the resultant normalized HOG descriptor feature vector.
- Legendre Moments and Local Binary Pattern based feature extraction
- Legendre moments for image of order p + q with intensity function f(x) is defined by the

$$L_{pq}(t) = \frac{(2p+1)(2q+1)}{4} \int_{-1}^{1} \int_{-1}^{1} P_p(x) P_q(y) f(x,y) d_x d_y \quad (5)$$

where Pp(x) is pth order legendre polynomials defined as.

$$P_p^{\tau}(t) = \sum_{k=0}^{i} \beth_{i,k} t^k, i = 0,1,2 \quad (6)$$

Where

$$\beth_{i,k} = \frac{(-1)^{i+k}(i+k)!}{(i-k)!\,7^k (k!)^2} \quad (7)$$

Legendre polynomials are orthogonal in nature and orthogonality condition is.

$$\int_0^1 P_i(t) P_j(t) dx = \begin{Bmatrix} 0 & if\ p \neq q \\ \frac{2}{2p+1} & otherwise \end{Bmatrix} \quad (8)$$

Using equation 8, any function can be written as.

$$f(x) = \sum_{i=0}^{m} c_i p_i(t) \quad (9)$$

Where

$$c_i = (2l+1) \int_0^1 f(t) P_i(t) dt \quad (10)$$

As $l \rightarrow \infty$ the sum of these functions is equal to the exact function. Basically, these character images are are 2D array of numerical values. So, first will introduce the mechanism for developing 2D Legendre polynomials. 2D Legendre moments can be generated by taking the product of two basic sets. Assume *Bm*(*x*), *Bm*(*t*) be the basis set of Legendre polynomials in two different variables. We can define the basis set of 2D images as *Bm*(*x, t*)=*Bm*(*x*)×*Bm*(*t*). The general term of the basis set *Bm*(*x, t*) can be written as

$$f(x) = \sum_{i=0}^{m} c_i p_i(t) \quad (11)$$

If we are using scale level *m*, then there must be (*m* + 1)2 terms in the basis set. Any images in the space *C*([0, 1] × [0, 1]) can be written as in term of Legendre polynomials as:

$$f(x,t) = \sum_{p=0}^{n} \sum_{q=0}^{m} c_{pq} P_p(x) P_q(t) \quad (12)$$

and coefficients can be obtained by the following relation as:

$$cpp = (2p+1)(2q+1) \int_0^1 \int_0^1 f(t,x) Pp(t) Pq(x) dt dx. \quad (13)$$

In hybrid method, we have used the combination of features of Linear binary pattern feature descriptor and Legendre moments based extracted features. These features are combined for further classification.

IV. CLASSIFICATIONS

After feature extraction phase, a technique is needed for the classification of these characters. Based on these extracted features, the classifier attempt to identify the pattern that represents the input character. There are many classifiers which makes the decision and can be divided into three types: structural, statistical and neural network classifiers. For instance, Support Vector Machines (SVM) [9] and Sequential Minimal Optimization algorithm for training Support Vector Machine are used. The classifiers take the extracted features as inputs. And classify other instances according to their properties. There are a large number of classifiers available for classification purposes, but in this article two widely used classifier i.e., A Sequential Minimal Optimization algorithm for training Support Vector Machine and Support Vector Machine are used in our experiments. A Sequential Minimal Optimization for training Support Vector Machine (SMO(SVM)) Sequential minimal optimization (SMO) is an algorithm for solving the quadratic programming (QP) problem that arises during the training of support vector machines. In binary classification problem, let s(x1, y1), ..., (xn, yn) be a datasets of sample and xi is an input vector and yiϵ−1, +1 is a binary label corresponding to each samples. Support vector machine (SVM) is trained by solving a quadratic programming problem, which is expressed in the dual form. Sequential Minimal Optimization (SMO) is an iterative algorithm for solving the optimization problem. SMO breaks a large problem into smallest possible sub-problems, which can be solved easily analytically. Because of the linear equality constraint involving the Lagrange multipliers.

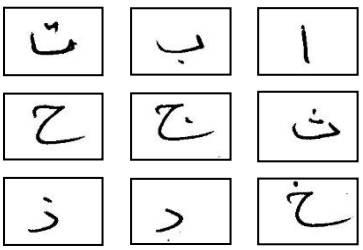

Figure 2. Sample of Arabic Characters in our Experiments the smallest possible problem involves two such multipliers. Then, for any two multipliers a1 and a2, the constraints are reduced to:

$$0 \leq \alpha_1, \alpha_2 \leq C,$$
$$y_1 \alpha_1 + y_2 \alpha_2 = k, \quad (14)$$

ISSN: 2409-2770

and this reduced problem can be easily solved analytically: we need to find a minimum of a one-dimensional quadratic function. K is the negative of the sum over the rest of the terms in the equality constraint, which is fixed in each iteration.

Support vector machines (SVM) are statistical models that separate two data sets [21]. The classes are divided through an optimal separating hyperplane (OSH). The limits of these classes of data and the OSH are called support vectors. Intuitively, for a set of points divided into two classes, the SVM method finds on one side the hyperplane that separates the highest possible fraction of points that belong to the same class, and on the other side it maximizes the distance between the classes and hyper plane. By using SVM, the user can avoid over-fitting due to its regularization parameter. The real help the SVM comes in is when there is randomly distributed data, taking into account a good training.

V. Experimental Setups and Results

We will compare the classification performance of three different feature extraction techniques. Our data sets are composed of isolated handwritten Arabic characters. The handwritten images are converted to gray-scale and normalized to a fixed-size image. In these experiments, data sets of each characters are manually created. A group of 30 people wrote each character on a plain paper. Every person in the group write each Arabic character 5 times. These words are scanned with the help of a scanner. The scanned documents are treated as single character.

Thus data set of 4004 isolated hand written Arabic characters are obtained. The resolutions of the dataset used in these experiments are 50 x 50 pixels. 63 % of the data are used to for training purpose and 37 % are used to check the classification performance. The sample of these images after segmentation are Shown in Figure 2. The implementation of each feature extraction techniques is designed in using MATLAB 2016.

A. Experimental Evaluation of Histogram Of Gradient Descriptor:

The results derived from confusion matrix shows that the classification performance of HOG descriptor in term of accuracy, Sequential Minimal Optimization has performed over the Support Vector Machine. As shown in Table 1, classification accuracy of 88.86% is obtained using Sequential Minimal Optimization technique for training support vector machine. The classifies correctly classify 1207 characters and incorrect classification rate is 154 characters. Other classification parameters are displayed in Table 1. The same extracted features are evaluated on Support Vector Machine. Accuracy of 16.09 % is achieved by using SVM as a classifier. As shown Table 1 total number of correctly classified instances are 219 and incorrectly classified instance are 1134. Detail Accuracy of both the classifiers i.e., SMO and SVM are given in Table 1.

Table 1: Histogram of Gradient Features Descriptor using SMO(SVM) and SVM

Summary of Histogram of Gradient Descriptor and Classifiers				
LBP using SMO(SVM)			*BP using (SVM)*	
Correctly classified Instance	1252	91.99%	1188	87.88%
Incorrectly classified Instance	109	8.008%	173	12.31%
Kappa Statistic	0.9169		0.8682	
Mean Absolute Error	0.00664		0.0091	
Relative Mean Square Error	96.32%		89.24%	
Relative Absolute Error	96.89%		13.18%	
Total Number of Instances	1361		1361	

Table 2: Local Binary Pattern Descriptor using SMO(SVM) and SVM

HOG using SMO(SVM) HOG using (SVM)			*HOG using SMO(SVM) HOG using (SVM)*	
Correctly classified Instance	1207	88.68%	219	16.09%
Incorrectly classified Instance	154	11.35%	1134	83.90%
Kappa Statistic	0.8862		0.134	
Mean Absolute Error	0.0664		0.0599	
Relative Mean Square Error	0.1799		0.2448	
Relative Absolute Error	96.34%		87.00%	
Total Number of Instances	1361		1361	

B. Experimental Evaluation of Hybrid Method

The results derived from confusion matrix shows that the classification performance of Hybrid methods in term of accuracy, Sequential Minimal Optimization has performed over the Support Vector Machine. As shown in Table 3, classification accuracy of 93.12% is obtained using Sequential Minimal Optimization technique for training support vector machine. The classifies correctly classify 1262 characters and incorrect classify 99 characters. Other classification parameters are displayed in Table 3. The same extracted features are also evaluated on Support Vector Machine. Accuracy of 88.48 % is achieved by using SVM as a classifier. As shown Table 3 total number of correctly classified instances are 1188 and incorrectly classified instance are 173. Detail Accuracy of both the classifiers i.e., SMO and SVM are given in Table 3.

ISSN: 2409-2770

Tabel 3: Hybrid feasture using SMO(SVM) and (SVM)

Hybrid features using SMO(SVM)			*Hybrid features using (SVM)*	
Correctly classified Instance	1262	93.12%	1188	88.48%
Incorrectly classified Instance	99	99	173	11.51%
Kappa Statistic	0.9169		0.8682	
Mean Absolute Error	0.0664		0.0089	
Relative Mean Square Error	79.99%		0.0089	
Relative Absolute Error	96.34%		51.64%	
Total Number of Instances	1361		51.64%	

VI. Comparison of Different Methods

In this section classification performance of three different feature extraction method is presented. In Table 4 the classification performance of these methods are shown. It is clear from Table 4 that for combination of extracted features using Legendre based features and LBP features along with Sequential Minimal Optimization algorithm for training SVM performs better then all other techniques. Classification Accuracy of 93.12% is achieved for SMO(SVM) in this experiment.

Table 4: Comparison of the all methods

		Correctly Classified Instances%	**Incorrectly Classified Instance%**
Hybrid Method	SMO	93.12	11.51
	SVM	88.48	6.88
Histogram of Gradient Descriptor	SMO	88.6	83.90
	SVM	83.90	11.31
Local Binary Feature Descriptor	SMO	12.711	12.711
	SVM	87.28	8.088
Legendre Based Features	SMO	30.34	86.336
	SVM	13.66	30.34

Conclusion

Main objectives of the paper is to find optimal combination of feature extraction methods and classifiers for the Classification of Arabic characters. In this paper we have investigated three different features extraction techniques with two classifiers namely SMO(SVM) and SVM. Result concluded from Section 6 that, best classification accuracy of 93.13% can be achieved when the combination of hybrid methods and sequential minimal algorithm for training support vector machine is used. Regard these all it can be that SMO(SVM) outperform better then SVM in all cases. In future works, finding the most accurate combination of feature extraction technique and classifier having more classification accuracy i.e, 100%.

References

[1] Fabrizio Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[2] Peter Jackson and Isabelle Moulinier, "Natural language processing for online applications:Text retrieval, extraction and categorization," volume 5. John Benjamins Publishing, 2007

[3] AM Mesleh, " Support vector machine text classifier for arabic articles: Ant colony optimization-based feature subset selection, "*The Arab Academy for Banking and Financial Sciences*, 2008.

[4] Franca Debole and Fabrizio Sebastiani, "An analysis of the relative hardness of reuters-21578 subsets, " *Journal of the Association for Information Science and Technology*,56(6):584–596, 2005.

[5] Abdelwadood Mohd Mesleh, "Support vector machines based arabic language text classification system: feature selection comparative study," In *Advances in Computer and Information Sciences and Engineering*, pages 11–16.Springer, 2008.

[6] Alaa M El-Halees.Arabic text classification using maximum entropy."*IUG Journa of Natural Studies*, 15(1), 2015.

[7] Mostafa M Syiam, Zaki T Fayed, and Mena B Habib, "An intelligent system for arabic text Categorization," *International Journal of Intelligent Computing and Information Sciences*, 6(1):1–19, 2006.

[8] Mohammad S Khorsheed and Abdulmohsen O Al-Thubaity," Comparative evaluation of text classification techniques using a large diverse arabic dataset, " *Language resources and evaluation*, 47(2):513–538, 2013.

[9] Bassam Al-Shargabi, Waseem Al-Romimah, and Fekry Olayah, "A comparative study for arabic text classification algorithms based on stop words elimination, " In *Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications*, page 11. ACM, 2011.

[10] Lambert Schomaker, Katrin Franke, and Marius Bulacu, "Using codebooks of fragmented. *Pattern Recog-nition Letters*, " 28(6):719–727, 2007.

[11] Horst Bunke and Kaspar Riesen, "Recent advances in graph-based pattern recognition with applications in document analysis, " *Pattern Recognition*, 44(5):1057–1067, 2011.

[12] G¨osta H Granlund, "Fourier preprocessing for hand print character recognition," *IEEE transactions on computers*, 100(2):195–201, 1972.

[13] H Al-Yousefi and SS Udpa, "Recognition of arabic characters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):853–857, 1992.

[14] K Roy, A Banerjee, and U Pal, "A system for word-wise handwritten script identification for indian postal automation, " In *India Annual Conference, 2004. Proceedings of the IEEE INDICON 2004. First*, pages 266–271. IEEE, 2004.

[15] Timo Ojala, Matti Pietik¨ainen, and David Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, 29(1):51–59, 1996.

[16] [16] Timo Ahonen, Abdenour Hadid, and Matti Pietik¨ainen, "Face recognition with local binary Patterns," *Computer vision-eccv 2004*, pages 469–481, 2004.

[17] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," In *Computer Vision and Pattern Recognition, 2005. CVPR2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE,2005.

[18] Seung Eun Lee, Kyungwon Min, and Taeweon Suh, "Accelerating histograms of oriented gradients descriptor extraction for pedestrian recognition, " *Computers & Electrical Engineering*, 39(4):1043–1048, 2013.

[19] Oscar D´eniz, Gloria Bueno, Jes´us Salido, and Fernando De la Torre, " Face recognition using histograms of oriented gradients, " *Pattern Recognition Letters*, 32(12):1598–1603,2011.

[20] Jon Arr´ospide, Luis Salgado, and Massimo Camplani," Image-based on-road vehicle detection using cost-effective histograms of oriented gradients, " *Journal of Visual Communication and Image Representation*, 24(7):1182–1190, 2013.

[21] Samir Al-Emami and Mike Usher, "On-line recognition of handwritten arabic characters,"*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):704–710, 1990.

Sungin Behram Khan belongs to Thana, Malakand, KPK, Pakistan. He has completed his bachelors in Electrical Engineering from UET Peshawar. He is currently the student of MS Electrical Engineering in UET Peshawar, Pakistan. His topic of research is pattern recognition and artificial intelligence.