Improved Arabic Words Classification using Spatial Pyramid Matching

Hassan Dawood,Hussain Dawood

Image Processing and Pattern Recognition Laboratory

Beijing Normal University

Beijing, China

hasandawod@yahoo.com,hussaindawood2002@yahoo.com

Ping Guo

Image Processing and Pattern Recognition Laboratory

Beijing Normal University

Beijing, China

pguo@ieee.org

Abstract—**In recent years, rapid abundance of hand written word classification attracted researchers to do Arabic word classification. Arabic language has cursive style of writing so it needs special framework for classification. In this paper, we present a precise framework for Arabic word classification using sparse coding with spatial pyramid matching (SPM) and linear support vector machines. SPM map each feature set to multiresolution histogram that preserves the individual features' at the finest level. The histogram pyramids are then compared using a weighted histogram intersection computation. We tested our proposed framework with four publically available datasets; IFN/ENIT, PATS-A01, IFHCDB and ISI Bangla numeral. Experimental results show that our proposed framework outperformed the state of art methods used for Arabic words classification.**

*Keywords-component; **Arabic Character Recognition, linear support vector machine (LSVM), Spatial Pyramid Matching (SPM), Scale invariant feature transform (SIFT).***

I. INTRODUCTION

From last two decades there has been lots of meaningful research work in pattern recognition for Offline word Recognition (OWR). There is much work done on Latin, Japanese, English and Chinese word classification. However, Arabic word images get less attention for most of researchers. They have specific features which make them different from languages over which classification is performed extensively in literature[24]. The specific features are (1) cursive nature in both machine-printed and hand-written text, (2) written from right to left instead of left to right and (3) the size in both height and width is not same. Arabic words recognition is still a challenging research area in pattern recognition. Arabic language partially used as script for several languages, like Farsi, Urdu, Uygur, Sindhi, Jawi, Kashmiri, Pishtu, Baluchi, Ottoman, Old Hausa, Berber, Dargwa, Ingush, Kazakh, and Kirghiz. It contains 28 characters from which 16 characters contain dots (dots vary from 1 to 3, which can be on upper side or lower side of character), combination of these characters make words. Some words look like same however diacritics make the different meanings.

Classification of Arabic words depends upon the following factors: (1) quality of image database; (2) effective and efficient preprocessing; (3) feature selection; and (4) classification method. Once the sample image is acquired, pre-processing is essential to improve the quality of image. Pre-processing includes thresholding, skew/slant correction, noise removal, thinning, baseline estimation and segmentation of words. After preprocessing, features are extracted from the preprocessed data and then classified. Arabic scripts can be recognized by two methods named as segmentation based and segmentation free. For the segmentation based method, the words are further divided into characters then used for recognition; this method is also known as analytical approach. For segmentation free methods, which are also known as a global approach, whole Arabic script image is given as an input for the Handwriting recognition (HWR) system. By using global approach the recognition process is made simpler by avoiding character segmentation, however it makes the recognition process slow due to huge vocabulary to process [1].

The main contribution of this paper is proposal of novel framework to classify the offline Arabic words using the spatial pyramid matching for both handwritten and computer typed Arabic words. Firstly, the handwritten and computer typed Arabic words are preprocessed. After the preprocessing the word descriptors are extracted using the scale invariant feature transform (SIFT), then sparse coding is applied, Spatial pyramid matching is used for the next step. For the sake of classification the linear support vector machine is adopted.

The rest of paper is organized into four sections. Section II describes related work, section III gives the overview to background knowledge, section IV describes our proposed algorithm, section V presents the detailed experiments and finally section VI provides conclusions and future work.

II. RELATED WORK

All over the world Arabic is written by more than 250 million people [2][1]. In early days people used their own (small and self generated) database [2] [3] so there is no way to compare their results for character classification. A recognition system is proposed based on a semi-continuous 1-dimensional Hidden Markov Model (HMM) [4]. Basic dimensions (like Height, length, and baseline skew) are normalized, features were extracted using a sliding window, and each binary word image normalization parameters were estimated [4]. Sabri et al. proposed Fast Hartley Transform (FHT) for feature selection, in which the contour of Arabic character is extracted and then the FHT is applied on extracted contour [5]. Hamdi et al. proposed Gabor filters for feature extraction with the KNN classifier and got better results than previously published approaches, Modified Fourier Spectrum [6] and Fourier descriptors are proposed [7] using the same

data [8]. Hamdi et al. proposed an algorithm for smoothing and segmenting the Arabic characters using width writing estimation from skeleton character. In it, moments, fourier descriptor of projection and centroids distance are taken as features of each character [9].

Mohannad et al. proposed an improved and more robust recognition algorithm for characters at the sub-word level based on Hidden Markov Models (HMMs) combined with the Viterbi algorithm. In it, statistical information is extracted and used to estimate the probabilistic parameters of HMMs [10]. Morteza et al. Used the SIFT features without any kind of preprocessing to remove the noise are used for word classification [11]. Mehmmood et al. proposed efficient feature extraction method in which features consist of the first three moment invariants, the number and position of dots, the number of holes in the character body, and the number of pixels in the dot area [12]. Jawad et al., after preprocessing extracted features extracted by using Discrete cosine Transform (DCT) and Zigzag scanning, and Neural Network (NN) based classifier is applied for word classification [13].

Spatial pyramid matching (SPM) has been shown to be effective for image classification. SPM segments the images using different scales and computes the Bag of words (BoW) histogram in each segment and then concatenates all the histograms [14]. Yang et al. proposed an approach to reduce the training complexity, a linear spatial pyramid matching method using sparse coding (ScSPM) is proposed [15]. Wang et al. used locality in feature space to constrain the linear sparse coding phase (LLC) of ScSPM, which has further reduced the computation time. However, the performance improvement of LLC over ScSPM on real world images is not obvious [16]. To the best of our knowledge, sparse coding with SPM is not used for Arabic word classification, it is first time to use in this paper.

III. BACKGROUND KNOWLEDGE

In this section, we review the previous works related to this word classification. Section 2.1 overviews SIFT. Section 2.2 gives the overview to sparse coding. Section 2.3 describes basics of Linear SPM and section 2.4 describes LSVM.

A. SIFT

Scale-invariant feature transform (SIFT) [23] is a common algorithm which detects and describes local features of images, and is applied vastly in computer vision. The SIFT algorithm has four filtering stages named as scale-space extrema detection, keypoint localization, orientation assignment and keypoint descriptor.

B. Sparse Coding

In this part, we will review the sparse coding [15]. Let us consider X (D- dimensional feature space) be a set of SIFT descriptors. i.e.

$$\min_{V} \sum_{m=1}^{M} \min_{K=1\ldots K} \|x_m - V_k\|^2$$

The vector quantization (VQ) method applies the K-means clustering algorithm to solve the following problem.

$$\min_{V} \sum_{m=1}^{M} \min_{K=1\ldots K} \|x_m - V_k\|^2$$

Where $V=[v_1,\ldots,v_k]^T$ are the K cluster centers to be found, called codebook, and $\|\ \|$ denotes the L2-norm of vectors. The optimization problem can be re-formulated into a matrix factorization problem with cluster membership indicators $U=[u_1,\ldots,u_m]^T$

$$\min_{u,v} \sum_{m=1}^{M} \|x_m - u_m V\|^2 \tag{1}$$

After the optimization, the index of the only nonzero element in u_m indicates which cluster the vector x_m belongs. In the training phase of VQ, the optimization Eq. (2) is solved with respect to both U and V. In the coding phase, the learned V will be applied for a new set of X and Eq. (1) will be solved with respect to U only. Then the VQ formulation is turned into another problem known as sparse coding (SC):

$$\min_{U,V} \sum_{m=1}^{M} \|x_m - u_m V\|^2 + \lambda|u_m| \tag{2}$$

where a unit L2-norm constraint on V_k is typically applied to avoid trivial solutions. Normally, the codebook V is an over complete basis set i.e. K>D dropped out the non-negativity constraint as well, because the sign of u is not essential—it can be easily absorbed by letting $V_m \leftarrow [V^T\,-V^T]$ and

$$u_m^T \ \leftarrow \ [u_{m+}^T \ - \ u_{m-}^T]$$

So that the constraint can be trivially satisfied, where $u_{m+}=$ min $(0,u_m)$ and $u_{m-}=$ max $(0,u_m)$. Similar to VQ, SC has a training phase and a coding phase. First, a descriptor set X from a random collection of image patches is used to solve Eq. (2) with respect to U and V, where V is retained as the codebook. In the coding phase, for each image represented as a descriptor set X, the SC codes are obtained by optimizing Eq. (2) with respect to U only.

C. Linear SPM

For any image represented by a set of descriptors, single feature vector can be computed based on some statistics of the descriptors' codes [15]. For example, if U is obtained via Eq. (1), a popular choice is to compute the histogram.

$$Z = \frac{1}{M} \sum_{m=1}^{M} u_m \tag{3}$$

The bag-of-words approach to image classification computes such a histogram z for each image I represented by an unordered set of local descriptors. In a more sophisticated SPM approach, the image's spatial pyramid histogram representation z is a concatenation of local histograms in various partitions of different scales. After normalization z can be seen as again a histogram. Let z_i denote the histogram

representation for image I_i. For a binary image classification problem, an SVM aims to learn a decision function

$$f(z) = \sum_{i=1}^{n} \alpha_i k(z, z_i) + b \qquad (4)$$

Where $\{(z_i, y_i)\}_{i=1}^{n}$ is the training set, and $y_i \in \{-1, +1\}$ indicates labels. For a test image represented by z, if f(z)>0 then the image is classified as positive, otherwise as negative. In theory k (.,.) can be any reasonable Mercer kernel function, but in practice the intersection kernel and Chi-square kernel have been found the most suitable on histogram representations. Linear kernel on histograms leads to always substantially worse results, partially due to the high quantization error of VQ. However, using these two nonlinear kernels, the SVM has to pay a high training cost, i.e. $O(n^3)$ in computation, and $O(n^2)$ in storage (for the n x n kernel matrix). This means that it is difficult to scale up the algorithm to the case where n is more than tens of thousands. Furthermore, as the number of support vectors scales linearly to the training size, the testing cost is O(n).

Let U be the result of applying the sparse coding Eq. (2) to a descriptor set X, assuming the codebook V to be pre-learned and fixed, image feature computed by a pre-chosen pooling function

$$z = F(U) \qquad (5)$$

where the pooling function F is defined on each column of U. Each column of U corresponds to the responses of all the local descriptors to one specific item in dictionary V. Therefore, different pooling functions construct different image statistics. For example, in Eq. (3), the underlying pooling function is defined as the averaging function, yielding the histogram feature. F is defined pooling as a max pooling function on the absolute sparse codes.

$$z_j = \max\left\{\left|u_{1j}\right|, \left|u_{2j}\right|, ..., \left|u_{Mj}\right|\right\}, \qquad (6)$$

Where z_j is the j-th element of z, u_{ij} is the matrix element at i-th row and j-th column of U, and M is the number of local descriptors in the region. This max pooling procedure is well established by biophysical evidence in visual cortex and is empirically justified by many algorithms applied to image categorization.

Similar to the construction of histograms in SPM, max pooling Eq. (6) used for spatial pyramid constructed for an image. By max pooling across different locations and over different spatial scales of the image, the pooled feature is more robust to local transformations than mean statistics in histogram. The pooled features from various locations and scales are then concatenated to form a spatial pyramid representation of the image. Let image I_i be represented by z_j, then simple linear SPM kernel is

$$k(z_i, z_j) = z_i^T z_j$$
$$z_i^T z_j = \sum_{l=0}^{2} \sum_{s=1}^{2^l} \sum_{t=1}^{2^l} \left\langle z_i^l(s,t), z_j^l(s,t) \right\rangle \qquad (7)$$

Where $\left\langle z_i, z_j \right\rangle = z_i^T z_j$ and $z_i^l(s,t)$ is the max pooling statistics of the descriptor sparse codes in the (s,t)-th segment of image I_i in the scale level l. Then the binary SVM decision function becomes

$$f(z) = \left(\sum_{i=1}^{n} \alpha_i z_i\right)^T z + b = w^T z + b \qquad (8)$$

In the literature, Eq. (4) is called the dual formulation of SVMs, while Eq. (8) is the primal formulation. Despite that the linear SPM kernel based on histograms leads to very poor performances, linear SPM kernel based on sparse coding statistics always achieves excellent classification accuracy. This success is largely due to three factors: (1) SC has much less quantization errors than VQ (2) it is well known that image patches are sparse in nature, and thus sparse coding is particularly suitable for image data and (3) the computed statistics by max pooling are more salient and robust to local translations.

D. Multi-class Linear SVM

Linear SVM is an algorithm for solving multiclass classification problems from very large data sets [15]. Linear SVM is a linearly scalable practice meaning that it produces an SVM model in a CPU time which scales linearly with the size of the training data set. Given the training data $\{(z_i, y_i)\}_{i=1}^{n}, y_i \in Y = \{1,..., L\}$, a linear SVM aims to learn L linear functions $\{w_c^T z \mid c \in Y\}$, such that, for a test datum z, its class label is predicted by

$$y = \max_{c \in Y} w_c^T \qquad (9)$$

We take a one-against-all strategy to train L binary linear SVMs, each solving the following unconstraint convex optimization problem

$$\min_{w_c}\left\{ j(w_c) = \|w_c\|^2 + c\sum_{i=1}^{n} l(w_c; y_i^c, z_i) \right\} \qquad (10)$$

Where $y_i^c = 1$ if $y_i = c$ otherwise $y_i^c = -1$, and $l(w_c; y_i^c, z_i)$ is a hinge loss function. The standard hinge loss function is not differentiable everywhere, which hampers the use of gradient-based optimization methods. We adopt a differentiable quadratic hinge loss,

$$l(w_c; y_i^c, z_i) = \left[\max\left(0, w_c^T z . y_i^c - 1\right)\right]^2.$$

Such that the training can be easily done with simple gradient-based optimization methods.
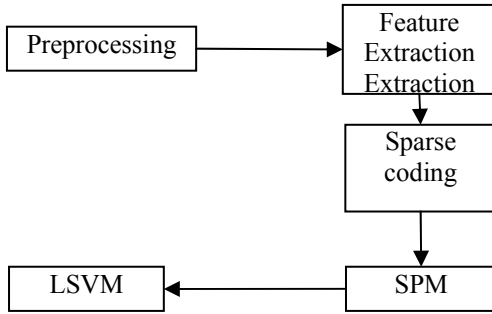
IV. PROPOSED ALGORITHM

Figure 1. Block diagram of the Framwork

Figure. 1 shows our proposed framework which has five main steps, described as following (1) Image binarization, smoothing, noise removal, word thinning, segmentation (2) Feature extraction, (3) Sparse coding, (4) Spatial pyramid matching and (5) Linear Support vector machine. The first phase provides preprocessing steps, however it is not compulsory that all preprocessing steps should be applied on the datasets. That can be determined by finding the threshold or visually give the quick overview to datasets. One dataset required all pre processing steps (like A01-Arial) however for other datasets there is no need of segmentation.

Form Figure 1 we can see that the first phase of framework is preprocessing. The question is, why preprocessing is needed: we can see from Figure 2, 4 and 5 the original images have some deficiencies that should be removed before extracting features .i.e. the original image in Figure has noise that should be removed to get the best possible features. For the image binarization, we have used a well known Otsu Method [17]. For the purpose of smoothing and noise removal method given in [18] is used, that exploits the fact that curves moving under their curvature, smooth out and disappear. This method evolves contours; boundaries remain essentially sharp and have no blur. For the purpose of thinning, we have used method given in [18]. The results of preprocessing step will be discussed in section 4.

 An image contains many objects and every object has many features. Interesting points on the object extracted as features of that object. The extracted descriptors can be used to locate the other objects in the same image.[23] SIFT image features provide a set of features of an object that are not affected by many of the complications experienced in other methods, such as object scaling and rotation. Important characteristic of these features is the relative positions between them in the original scene should not change from one image to another. SIFT detects and uses a much larger number of features from the images, which reduces the contribution of the errors caused by these local variations in the average error of all feature matching errors. SIFT features are also very resilient to the effects of "noise" in the image

The Bag of Feature (BoF) method removes the spatial sequence of local descriptors, which restricts the descriptive power for image representation. By solving this problem, one particular expansion of the BoF model, called SPM [14] has made an outstanding success in image classification.SPM mechanism involves the segmentation of image in to increasing fine sub-regions and calculate the histogram of local features originate inside the sub-regions. It improves the performance over basic bag-of-features representation also geometric based methods. A pyramid match kernel works with an order less image representation. It permits for precise matching of two groups of features in a high dimensional appearance space, but discards all spatial information.

Linear SVM maximizes the geometric margin of training dataset. Linear SVM is algorithm for solving multiclass classification problems for very large datasets. Linear SVM is a linearly scalable practice meaning that it produces an SVM model in a CPU time which scales linearly with the size of the training data set. It is Binary and linearly separable classification.

V. EXPERIMENTS

We have used the bench mark datasets for classification which are IFN/ENIT, PATS-A01, IFHCDB and ISI Bangla numeral.

A. Datasets

The first Printed Arabic Text Set A01 (PATS-A01) consists of 2766 text line images. The line images are available in eight fonts: Arial, Tahoma, Akhbar, Thuluth, Naskh, Simplified Arabic, Andalus, and Traditional Arabic. We used only arial font in our experiments. Text lines are segmented to words in preprocessing step. We used 70% images for training dataset and remaining images as testing data [19].

The IFN/ENIT-database contains 32,493 binary word images with more than 2200 classes and written by 411 writers. We used 70% images for training data set and 30% images for testing data [4] [13] [10].The IFHCDB Farsi database has 69,855 numeral images in total. It contains 47 classes. We used 70% images for training data set and rest images for testing data [20].

The ISI Bangla numeral database has 19,392 training samples and 4000 test samples in total. The images are gray-scaled, some with noisy background, and the gray level of foreground varies considerably. We used 12000 images for training data set and 3000 images for testing data [21].

B. Performance Measure

Performance of existing and our proposed framework is compared in terms of accuracy. Accuracy is calculated in term of the ratio of real correct images to images annotated by human. Table 1, shows that how our proposed framework gives more accurate results.

C. Baseline Approaches

We compared our proposed framework with work presented in [11] which tested framework on 1400 text images and claims that there is no need of preprocessing. However, in our case the dataset size is much larger and needs to be preprocessed. AIFN/ENIT dataset is used and good results are obtained by [22]. However our proposed framework also out performs this framework.

D. Results and Discussions

For [19] the images are segmented into words in preprocessing (all steps of preprocessing defined in introduction) steps the example of preprocessing for this data set shown in Figure 2 and 3.

Figure 2. PATS-A01 Original image [19]

Figure 3. Segmented image after Preprocessing [19]

For the dataset Bangla numeral images have noise due to noise these images were not clear so we have applied some smoothing and noise reduction techniques [18] to made the dataset ready for feature extraction by SIFT. Some examples are shown in Figure 4 before and after preprocessing.
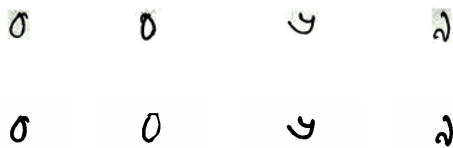

Figure 4. Bangla numeral dataset first row before preprocessing 2nd row after the noise removal [BAN].

For the IFN/ENIT [RAF 04] the images were not sharp so for making the images sharp, firstly, we find the edges then function imfill is used to sharp the image. Before and after sharpening the images examples are shown in Figure 5.
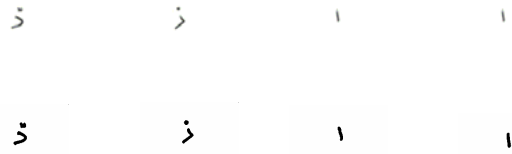

Figure 5. IFN/ENIT dataset first row before preprocessing 2nd row after image enhancement [MOH 11].

Experimental results are shown in Table 1. It is clear that for all four standard benchmark datasets our framework out performs baselines. We have compared our framework with the DCT and SIFT using the SVM. For the preprocessing we have used the different criteria for different data sets according to requirements of datasets.

Method /Data set	DCT + SVM	SIFT + SVM	Proposed Framework
PATS-A01	98%	98.8%	100%
IFN/ENIT	45%	71%	100%
Bangla	30%	75%	93%
IFHCDB	47%	70%	88%

Table 1. Classification Results

PATS-A01 is computer written words and has same diacritics, so the accuracy is good and almost same for all methods.

IFHCDB is handwritten words and much variation in diacritics also has different type of noises in it, so extracted features are unordered and baseline methods are unable to cater relationship between words. Meanwhile SPM has the capability to cater the unordered feature.

IFN/ENIT and Bangla is handwritten characters, contains noises and not sharp. So base line approaches did not performed on it and our framework out performs on it because we have done preprocessing for removal of noise make it sharpen. SPM map each feature set to multiresolution histogram that preserves the individual features' at the finest level. The histogram pyramids are then compared using a weighted histogram intersection computation, so we got the more accurate results on datasets.

VI. CONCLUSIONS

This paper proposed a method to classify the Arabic hand written as well as Computer generated words using SPM. The features are extracted by SIFT and our proposed algorithm is evaluated with four datasets. Sparse coding is used for

dictionary learning. Our proposed algorithm has shown promising results which produces high rate of classification. LSVM as classification tool in Arabic character recognition system got promising results. For the future work it will be considered that how to automatically find which datasets requires which preprocessing steps.

REFERENCES

[1] M.S. Khorsheed.," Off-Line Arabic Character Recognition - A Review", PAA. vol.5, 31-45 (2002).

[2] Adnan, A.," Off-line Arabic character recognition: The state of the art". JPR, 31, 517–530 (1998).

[3] Badr, A. B., Sabri A. M..," Survey and bibliography of Arabic optical text recognition", SP.41(1), 49– 77 (1995).

[4] Mario. Pechwitz , Volker. Maergner.,"HMM-based approach for handwritten Arabic word recognition using the IFN/ENIT − database". ICDAR, 890–894 (2003).

[5] Sabri A. Mahmoud , Ashraf S. Mahmoud,"The use of Hartley transform in OCR with application to printed Arabic character recognition". PAA . 12, 353-365 (2009).

[6] Mahmoud S.A., Mahmoud A.S.,"Recognition using Modified Fourier Spectrum (MFS)".*PICGMI,* 155 – 159 (2006).

[7] Sabri .A. Mahmoud.,"Arabic character recognition using fourier descriptors and character contour encoding". PR. 27, 815-824(1994).

[8] Hamdi A. Al-Jamimi , Sabri A. Mahmoud," Arabic Character Recognition Using Gabor Filters". Innovations and Advances in Computer Science and engineering , 113-118 (2010).

[9] Hamdi, R, Bouchareb, F. Bedda, M," Handwritten Arabic character recognition based on SVM Classifier". ICTTA , 1-4 .(2008).

[10] Mohannad M. Aulama, Asem M. Natsheh, Gheith A. Abandah., "Optical character recognition of handwritten Arabic using hidden Markov models ",Proc. SPIE 8055, 80550L (2011).

[11] Morteza Zahedia ,Saeideh Eslami," Farsi/Arabic optical font recognition using SIFT features". PCS,3,1055-1059 (2011).

[12] Mehmmood Abd, George Paschos,"Effective Arabic Character Recognition using Support Vector Machines". IACS E , 7-11 (2007).

[13] AlKhateeb, J.H., Jinchang Ren, Jianmin Jiang, Ipson, S.S, El Abed, H.,"Word-based Handwritten Arabic Scripts Recognition using DCT Features and Neural Network Classifier". SSD. 1-5 (2008).

[14] Lazebnik S., Schmid C., Ponce J., "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories". CVPR, 2169 – 2178 (2006).

[15] Yang, J., Yu, K., Gong, Y., Huang, T.," Linear spatial pyramid matching using sparse coding for image classification". CVPR,(2009) .

[16] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.," Locality-constrained linear coding for image classification". CVPR ,(2010).

[17] Otsu N,"A threshold selection method from gray-level histogram". IEEE Trans. SMC, 62–66 (1979).

[18] Rafael C.Gonzalez,Richarh E.Woods Steven   L.Eddins.," Digital Image Processing Using MATLAB*"* .(2004)*.*

[19] http://www.ccse.kfupm.edu.sa/~husni/ArabicOCR/PATS-A01.htm

[20] Mozaffari S, K. Faez, F. Faradji, M. Ziaratban and S.M.Golzan".A comprehensive isolated Farsi/Arabic character database for handwritten OCR research". Proc. IWFHR, 385-389 (2006).

[21] http://www.isical.ac.in/~ujjwal/download/BanglaNumeral.html

[22] Jawad H. AlKhateeb, Jianmin Jiang, Jinchang Ren, Fouad Khelifi , Stan S. Ipson, "Multiclass Classification of Unconstrained Handwritten Arabic Words Using Machine Learning Approaches".2,.21-28 (2009).

[23] David G. Lowe," Distinctive Image Features from Scale-Invariant Keypoints" .IJCV.60, 91-110 (2004).

[24] L.Douglas Baker,A Kachites McCallum," Distributional Clustering of Words for Text Classification". ACM, (1998).