

RECOGNITION IMAGE OBTAINED BY CAMERA PHONE OF CHARACTER ARABIC

¹Youssef Rachidi and ²Zouhir Mahani

¹Laboratoire Image et Reconnaissance de Formes – Systèmes Intelligents et Communicants, (IRF – SIC),
Université Ibn Zohr, Agadir, Maroc

²Laboratoire Des Sciences de l'Ingénieur et Management de l'Energie, (LSIME), Université Ibn Zohr,
Agadir, Maroc

ARTICLE INFO

Article History:

Received 20th January, 2018
Received in revised form
06th February, 2018
Accepted 17th March, 2018
Published online 30th April, 2018

Key Words:

Pretreatments,
Arabic characters,
Mobile phone, OCR,
CNN, Random Forest.

ABSTRACT

In this paper, we proposed an offline Arabic handwriting character recognition system for isolated characters obtained by camera phone. Initially doing some pretreatments on the picture, the text is segmented into lines and then into characters. In a second phase, we have employed a several methods for extracting the features from the handwriting Arabic character, these methods are: Grey Level Co-occurrence Matrix (GLCM), Gabor Filters, Zoning, Projection Histogram, and Distance Profile. In addition, we have also tested the various combinations of Gabor Filters and Zoning. After that, for the classification stage we have used two classifiers: the Random Forest Method and Convolutional Neural Networks. However, we have presented a comparison between these classifiers. We carried out the experiments with a database containing 2800 samples collected from different writers. The experimental results show that our proposed OCR system is very efficient and provides good recognition accuracy rate of handwriting Arabic characters images acquired via camera phone.

Copyright © 2018, Youssef Rachidi and Zouhir Mahani. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Citation: Youssef Rachidi and Zouhir Mahani, 2018. "Recognition image obtained by camera phone of character arabic", *International Journal of Development Research*, 8, (04), 20147-20151.

INTRODUCTION

The automatic recognition of handwritten or printed Arabic characters remains a subject of research and experimentation. The problem is not yet solved despite the fact that results have reached fairly high rates in some applications (Svetnik, 2003). Some attempts have been done to improve the current situation. In this context, we have employed a recognition system of Arabic handwritten characters extracted from a picture taken by camera phone (Hassan El Bahi, 2015). Indeed, in the primitives' extraction stage, our approach is based on primitives of the Zoning types (Elima Hussain, 2015), of distance profile feature (Siddharth, 2011). Projection histogram and Gray Level Co-occurrence Matrix (GLCM) technique (Ggg, 1973). These primitives will supply a Random Forest Method and CNN in the learning and recognizing phases. On a database of handwritten, segmented and isolated characters acquired by camera phone, obtained an encouraging results on the majority of this characters.

*Corresponding author: Youssef Rachidi

Laboratoire Image et Reconnaissance de Formes – Systèmes Intelligents et Communicants, (IRF – SIC), Université Ibn Zohr, Agadir, Maroc

In this paper, our objective is mainly interested in the development of handwriting Arabic character recognition system and Improvement of the Recognition Rate by Random Forest and CNN, in which the images are obtained by camera phone.

Pre-Processing

The procedure of preprocessing which refines the scanned input image includes several steps: Binarization, for transforming gray-scale images in to black and white images, noises removal, and skew correction performed to align the input paper document with the coordinate system of the scanner and segmentation into isolated characters (Svetnik, 2003).

Binarization and Noise Remo- oval

We used the Sauvola method for (Sauvola, 200), this method of thresholding is performed as a preprocessing step to remove the background noise from the picture prior to extraction of characters and recognition of text. Fig.2 (a) shows a sample

input handwritten arabic character image and Fig.2(b) shows the binarized image after the thresholding step using Sauvola method. Noise which is in the images is one of the big difficulties in optical character recognition process. The aim of this part is to remove and eliminate this obstacle; there are several methods that allow us to overcome this problem. In this work we decided to use the morphology operations to detect and delete small areas of less than 30 pixels (Hassan El Bahi, 2015).

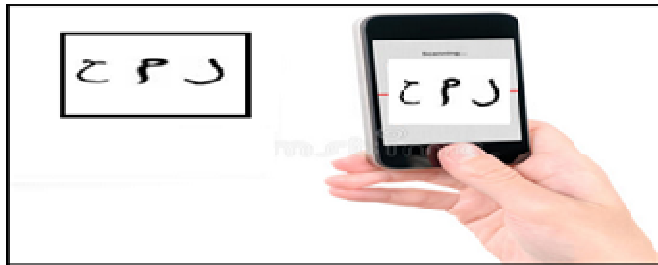


Fig. 1. Image acquisition

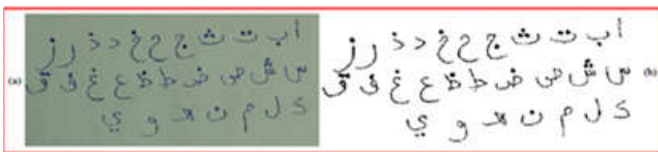


Fig. 2. (a) Example of an input image, (b) Thresholded image with Sauvola method

Skew detection and correction

Skew correction methods are used to align the paper document with the coordinate system of the scanner. Main approaches for skew detection include line correlation (Yan, 1993), projection profiles (Pavlidis, 1992). Hough transform (Le, 1994) etc. For this purpose two steps are applied. First, the skew angle is estimated. Second, the input image is rotated by the estimated skew angle. In this paper, we use the Hough transform to estimate a skew angle θ_s and to rotate the image by θ_s in the opposite direction.

Segmentation

Next step for OCR is the Segmentation of the image. In This paper we propose a segmentation algorithm, in which text is easily segmented into Lines and Words using the traditional vertical and horizontal projection (Archana, 2012).

Line Segmentation

Once the image of the text cleaned, the text is segmented into lines. This is used to divide text of document into individual lines for further preprocessing. For this, we used analysis techniques of horizontal projection histogram of the pixels in order to distinguish areas of high density (lines) of low-density areas (the spaces between the lines) (see Fig.3). These techniques were often used to extract lines in printed texts (Svetnik, 2003).

Characters Segmentation

We used in this part the vertical projection histogram to segment each text line of characters. Fig.4 shows a text line, the vertical histogram and the result of segmentation into characters (Hassan, 2015).

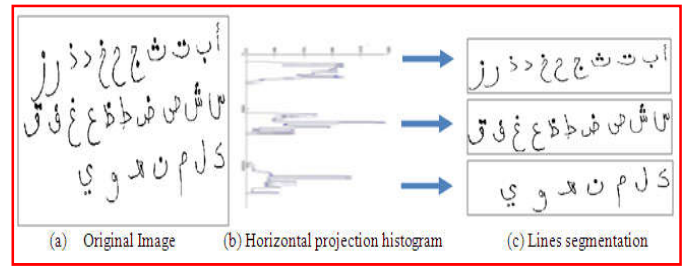


Fig. 3. Lines segmentation

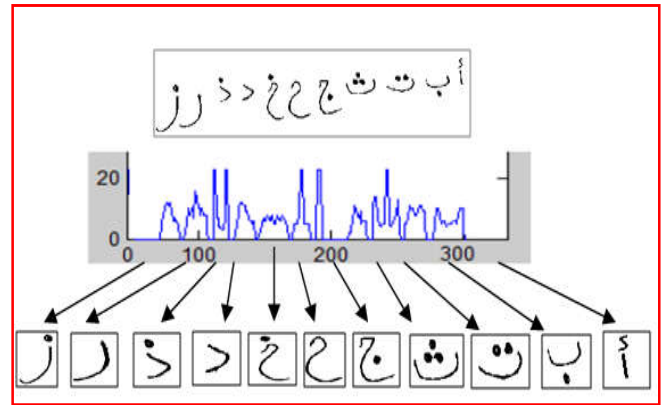


Fig. 4. Characters segmentation

In This part we present some feature extraction methods for recognition of segmented (isolated) characters (Ivind Due Trier, 1995). Selection of a feature extraction method is probably the single most important factor in achieving high recognition performance in character recognition systems. Different feature extraction methods are designed for different representations of the characters, such as solid binary characters, character contours, skeletons (thinned characters) or gray-level sub-images of each individual character (Ivind Due Trier, 1995). In this paper, we have tested four methods: the Zoning types, Distance profile feature, Projection histogram and Gray Level Co-occurrence Matrix (GLCM) technique.

Gabor filters

As a powerful feature, the Gabor filters (Daugman, 1980) have been successfully applied in numerous pattern recognitions including face recognition fingerprint recognition ..., as well as optical characters recognition. The Gabor filters are defined by a complex sinusoidal modulated by a Gaussian envelope described as follows:

$$G(x, y, \theta, f) = e^{-\frac{1}{2} \left(\frac{R_1^2}{\sigma_x^2} + \frac{R_2^2}{\sigma_y^2} \right)} \cos(2\pi f x_\theta)$$

Where:

$$R_1 = x \cos(\theta) + y \sin(\theta)$$

$$R_2 = y \cos(\theta) - x \sin(\theta)$$

f represents the frequency of the Sinusoidal plane wave along the direction θ , and (σ_x, σ_y) explain the standard deviations of the Gaussian envelope along x and y directions (Daugman, 1980).

Zoning

The zoning (Elima Hussain, 2015), is a statistical region-based feature extraction, it aim is to get the local characteristics in

lieu of global characteristic. Therefore, according to the size normalized character image (60 x 50 pixels), we divided it into 30 (6 x 5) zones of 10 x 10 pixels size, then we calculated the densities of pixels in each zone, finally we are getting 30 features.

Projection histogram

Projection histogram descriptor is a statistical feature, According to this feature we have used two direction of projection horizontal traversing. The horizontal histogram of the character arabic computed by counting the number of black pixels in each row. At the last we will have 60 features depending on the direction projection.

Gray Level Co-occurrence Matrix

Gray Level Co-occurrence Matrix (GLCM) technique is an approach for extracting statistical texture features that have been proposed by Haralick (Ggg, 1973). The main principle of GLCM is to counts the number of times various combinations of pixel gray levels occur in a given image. Haralick defines 14 statistical features measured from the GLCM. In this work, five important features are used namely energy, contrast, correlation, entropy and homogeneity.

Distance profile

In distance profile feature (Siddharth, 2011), the distance (number of pixels) between the bounding box of image and the first pixel of foreground will be calculated. We have employed two types of profiles sides left and top. Concerning left profile, it is extracted by counting the distance from the left bounding box to the nearest foreground pixels in each row. Then as well, top profile, it is extracted by counting the distance from the top bounding box to the nearest foreground pixels in each column.

Table 1. Combination of the different feature vectors

Feature Method	Contained feature	Size
FM1	GLCM	5
FM2	Gabor filters	32
FM3	Zoning	30
FM4	Projection Histogram Horizontal	60
FM5	Distance Profile (Left + Top)	110
FM6	Gabor filters + Zoning	62

Classification

In the complete process of system recognition of forms, the classification plays an important role by pronouncing on the membership of a shape in a class. The main idea of the classification is to attribute an example (A form) not known about one Class predefined from the description in parameters of the form. Several surrounding areas of classification are used in the field of recognition of forms which are more or less good adapted to the recognition of the writing. In litterateur, there are many types of classifiers that have been implemented in handwritten optical Arabic character recognition problems. Among them, in this paper we have used two classifiers: the Convolutif Neural Network (CNN) and Random Forest.

Convolutif Neural Network

Convolutional networks are derived from perceptron architectures Multi Layer Perceptron (MLP), however they use shared weights, related to the convolution window, which allow them an implicit extraction of local features. The difference of Convolutional neuron networks compared to conventional networks of MLP type, let us analyze the principle of recognition on the character (" R" RAE in Arabic), Fig.5 A neuron of an MLP is fully connected to all the neurons of the previous layer while for a convolutional network, a neuron is connected to a subset of neurons of the previous layer. Each neuron can be seen as a unit for detecting a local characteristic, a particular structural singularity such as the detection of a vertical or horizontal line, or even a loop. Along the trajectory, the matrix of weights corresponding to the sliding window is identical (notion of shared weights): same detection, same convolution

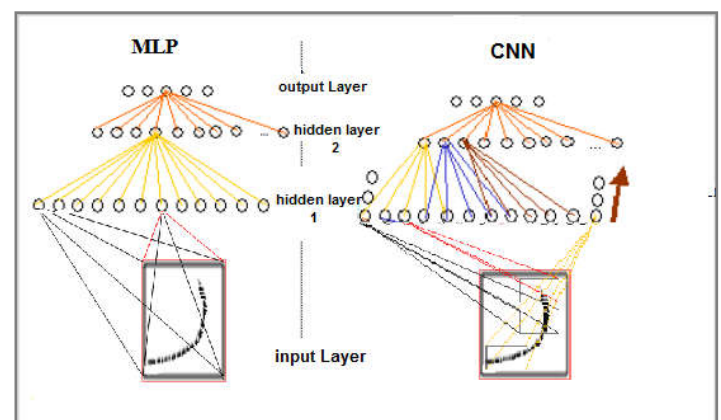


Fig. 5. The difference of Convolutional neuron networks compared to conventional networks of MLP type

Random Forest

Random forest is an ensemble training algorithm that constructs multiple decision trees. It suppresses over-fitting to the training samples by random selection of training samples for tree construction in the same way as is done in bagging (Breiman,1996) (Breiman,1999) resulting in construction of a classifier that is robust against noise. Also, random selection of features to be used at splitting nodes enables fast training, even if the dimensionality of the feature vector is large (Svetnik, 1958).

Algorithm

$z = \{(x_1, y_1), \dots, (x_n, y_n)\}$ learning sample, x_i describes nominal variables p explanatory [20]:

1. for $b=1$ to B (B number of trees)
 - (a) Draw a bootstrap sample z_b of size N from the training data
 - (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{\min} is reached.
 - i. Select m variables at random from the p variable
 - ii. Pick the variable/split-point among the m
 - iii. Split the node into two daughter nodes
2. Output the ensemble of tree $\{T_b\}_1^B$

To make a prediction at a new point x :
Regression:

$$\hat{f}_{rf}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x) \dots\dots\dots(4)$$

Classification: let $\hat{C}_b(x)$ be the class prediction of the b th random-forest tree. Then

$$\hat{C}_{rf}^B(x) = \text{majority vote } \{\hat{C}_b(x)\}_1^B \dots\dots\dots(5)$$

Why Random Forest works [20]

Mean Squared Error = Variance + Bias²

If trees are sufficiently deep, they have very small bias

How could we improve the variance?

$$\text{var} \left(\frac{1}{B} \sum_{i=1}^B T_i(c) \right) = \frac{1}{B^2} \sum_{i=1}^B \sum_{j=1}^B \text{Cov}(T_i(x), T_j(x)) \dots(6)$$

EXPERIMENTAL RESULTS






Due to the absence of standard database of handwritten Arabic characters acquired by camera phone, we have constructed our own database of upper-case Arabic character (“alif” to “alyae”) images obtained by camera phone. The database contains 100 samples of 28 classes, collected from 10 different writers. As a result the database consists of 2800 samples. The samples are divided randomly into two set, one for training stage, we have used 85 % (2380 samples) and the other for testing stage, we have used 15 % (420 samples). We have tested the proposed system on database of handwritten Arabic characters acquired by camera phone the SAMSUNG Galaxy N6 of this characteristics; 12 Megapixels. For classification stage we have used two classifiers: the Convolutif Neural Network (CNN) and the Random Forest and for each classifier we employed a set of different features extraction methods. The Zoning feature extraction (FM1) provides higher recognition and learning rate by Random Forest with the achievement of a Recognition rate of 93.54 %. Also FM 4 and FM 3 give some encouraging results. According to the results of Convolutif Neural Network the hybrid method FM 2 achieves a very good recognition and training rate: 95, 22 % of Learning rate and 93, 35 % of Recognition rate.

Table 2. Results of different single feature vectors using Convolutif Neural Network and Random Forest classifiers

Classifier Feature Vector	Random Forest (N=600)		Convolutif Neural Network	
	Learning R.	Recognition R.	Learning R.	Recognition R.
FM1	93,61%	91,32%	95,22%	93,35%
FM2	80,46%	78,25%	83,11%	82,31%
FM3	95,16%	93,54%	94,64%	92,11%
FM4	93,94%	93,46%	94,19%	91,73%
FM5	92,89%	92,71%	92,68%	88,24%
FM6	95,76%	94,84%	93,05%	93,42%

Out of the 420 Read-only characters, 392 were recognized, representing a recognition rate of 93.54%. With respect to the rate obtained for each letter, the best result achieved with this approach was 98.12%, for the character (Ha). Table 3 below shows the recognition rate obtained on certain characters.

Table 3. Below shows the recognition rate obtained on certain characters

Characters	Recognition Rate
 (AAIN)	73,58%
 (MIM)	93,43%
 (HA)	98,12%
 (KAF)	86,17%
 (RAE)	92,36%

The recognition errors are high for the letter “AAIN”, which is explained in particular by the insufficiency of the characteristics used to better describe each character during phase of extraction the primitives, and to the initial data used during the learning step. A good estimate of this data can reduce the error rate of our system.

Conclusion

In this paper, we have presented a system of handwriting Arabic character recognition based on the method Random Forest and Convolutif Neural Network. Several features have been studied and compared; as a result we’ve chosen Sauvola [10] method duo its ability to remove the noise. The experiments carried out in database were performed on a database obtained by camera phone with applying different classifiers and for each classifier we have tested a set of single feature methods. The results obtained in this paper that has been compared and analyzed have shown that Random Forest with Zoning feature is the best in terms of recognition accuracy rate and GLCM technique provide higher recognition rate by CNN. In future work, we will add other features methods that improve the results for some characters for example, minimize the length of execution of program which to calculate the recognition rate.

REFERENCES

Archana A. Shinde, D.G.Chougule, “Text Pre-processing and Text Segmentation for OCR” *IJCSET |January 2012| Vol 2, Issue 1,810-812*
 Breiman, L. 1996. Bagging predictors. In *Machine Learning*, Springer.
 Breiman, L. 1999. Using adaptive bagging to debias regressions. In *Technical Report*. Statistics Dept. UCB.
 D. S. Le, G. R.Thoma and H. Wechsler, “Automatic page orientation and skew angle detection for binary document images”, *Pattern Recognition* 27, 1994, 1325-1344.

- Daugman, J.G. 1980."Two-dimensional spectral analysis of cortical receptive field profile," *Vision Research*, 20, pp. 847-856, 1980.
- David Bouchain *Character Recognition Using Convolutional Neural Networks*. Seminar Statistical Learning Theory. University of Ulm, Germany. Institute for Neural Information Processing. winter 2006-2007
- Elima Hussain, Abdul Hannan, Kishore Kashyap, "A Zoning based Feature Extraction method for Recognition of Handwritten Assamese Characters" *IJCST Vol. 6, Issue 2*, April - June 2015
- Ggg R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural Features for Image Classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, pp. 610-621, 1973.
- H.Yan, "Skew correction of document images using interline cross-correlation", *CVGIP: Graphical Models Image Process* 55, 1993, 538-543
- Hassan El Bahi, Zouhir Mahani and Abdelkarim Zatni "A robust system for printed and handwritten character recognition of images obtained by camera phone" *.Published in WSEAS Transactions on Signal Processing, Volume 11*, 2015, pp. 9-22
- Ivind Due Trier, Anil K. Jain, Torfinn Taxt, "Feature extraction methods for character recognition a survey" Revised July 19, 1995
- Pavlidis, T. and J. Zhou, "Page segmentation and classification", *Comput. Vision Graphics Image Process.* 54, 1992, 484-496
- Sauvola, J. and M. Pietikainen, "Adaptive Document Image Binarization," *Pattern Recognition* 33(2), pp. 225-236, 2000
- Siddharth, K.S., Jangid, M., Dhir, R., Rani, R., "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features", *International Journal on Computer Science and Engineering (IJCSE)*, Vol. 3 No. 6, pp. 2332-2345, 2011
- Svetnik, A. Liaw, C. Tong, J. Culbertson, R. Sheridan, and B. Feuston. Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43:1947-1958, 2003.
