Arabic Character Recognition based on Statistical Features
A Comparative Study

Mariem Gargouri Kchaou[1], Slim Kanoun[1], Fouad Slimane[1,2] and Souhir Bouaziz Affes[1]

[1] *University of Sfax, National School of Engineers, BP 1173, Sfax, 3038, Tunisia*

[2] *DIVA Group, University of Fribourg, Bd de Pérolles 90, CH-1700 Fribourg, Switzerland*

Keywords: Arabic Optical Character Recognition, Statistic Approach, Features Extraction, Classification, Offline Recognition.

Abstract: This paper presents a comparative study for Arabic optical character recognition techniques according to statistic approach. So, the current work consists in experimenting character image characterization and matching to show the most robust and reliable techniques. For features extraction phase, we test invariant moments, affine moment invariants, Tsirikolias–Mertzios moments, Zernike moments, Fourier-Mellin transform and Fourier descriptors. And for the classification phase, we use k-Nearest Neighbors and Support Vector Machine. Our data collection encloses 3 datasets. The first contains 2320 multi-font and multi-scale printed samples. The second contains 9280 multi-font, multi-scale and multi-oriented printed samples. And, the third contains 2900 handwritten samples which are extracted from the IFN/ENIT data. The aim was to cover a wide spectrum of Arabic characters complexity. The best performance rates found for each dataset are 99.91%, 99.26% and 66.68% respectively.

1 INTRODUCTION

Today, in the entire world, most of the information is reserved and developed by computers. Nevertheless, information is still collected by using paper. Information in paper form is difficult to manipulate. So, it must to be changed to computer information. Optical Character Recognition (OCR) is an efficient method that a machine can extract and save the information automatically.

OCR systems are categorized into two domains. One focuses on picture detection of letters after entrance to system that is called offline recognition. In the other domain however, the writer enters the texts directly to system that is called online recognition. So in this study, we concentrate on characters that are collected in offline mode.

Character recognition is an attractive subject in the field of pattern recognition. Unlike English language, there has been only a few works on Arabic characters recognition (Abandah et al., 2009); (Abdul Sattar and Shah, 2012); (Imran et al., 2012); (Jenabzadeh et al., 2011); (Zaghloul et al., 2011). However, Arabic texts have main specifications which make them difficult to segment and to recognize. Arabic script is cursive in nature and the segmentation into characters is difficult in printing as well as in handwriting. An Arabic character might have several shape forms (1 to 4 shapes) depending on its relative position in the word. For example, the character Ain has four forms: isolated (ع), initial (عـ), medial (ـعـ) and final (ـع). In addition, some Arabic characters have the same shape and differ from each other only by existing of dots. For example, Jiim (ج) has a dot under its main body, Haaa (ح) hasn't dots, and Khaa (خ) has a dot above its main body.

OCR systems have four stages, that each of them has its own problems and effects on the system. These four stages are pre-processing; feature extracting, character categorization and post-processing.

In this paper, we propose a comparative assessment above feature extraction techniques and classifiers. We expose a comparison between the studied features and the used classifiers to draw conclusions. In the literature, many comparative studies (Aboaisha et al., 2012); (Mozaffari et al., 2004) exist, but at our knowledge, there is no study which has been addressed as follow: 3 datasets, 6 feature extraction techniques and 2 classifiers.

The remainder of this paper is organized as follows: We present in Section 2 the used features

Gargouri Kchaou M., Kanoun S., Slimane F. and Bouaziz Affes S. (2013).
Arabic Character Recognition based on Statistical Features - A Comparative Study.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 281-284
DOI: 10.5220/0004265602810284

and classifier. Section 3 details the experimental setup by using printed, multi-oriented and handwritten Arabic characters. Finally, the conclusions and the future works can be found in Section 4.

2 FEATURES AND CLASSIFIERS

The task of feature extraction is to reduce the data by measuring “features” or “properties” that distinguish between different characters. These features are then passed to a classifier that evaluates the evidence presented and makes a final decision. Feature selection and extraction plays an important role in pattern recognition.

The features extraction techniques are invariant moments, affine moment invariants, Tsirikolias–Mertzios moments, Zernike moments, Fourier-Mellin transform and Fourier descriptors. We also test some features combinations (some or all of them). Features vector size of:

- Hu’s Invariant Moments (IM) (Imran et al., 2012) is 7.
- Affine Moment Invariants (AMI) is 6.
- Tsirikolias–Mertzios Moments (TMM) is 12.
- Zernike Moments (ZM) (Abandah et al., 2009); (Aboaisha et al., 2012) is 20.
- Fourier-Mellin Transform (FMT) is 17.
- Fourier Descriptors (FD) (Sabri and Ashraf, 2009) is 9.

There is features which are not used before on Arabic characters recognition, like AMI, TMM and FMT.

The classifiers are k-Nearest Neighbors (k-NN) and Support Vector Machine (SVM).

We integrated the fuzzy dimension in k-NN classifier to improve our system with the intelligent aspect. We also studied distance-weighted k-NN. We tested traditional k-NN (T), weighted k-NN (W) and Fuzzy k-NN (F). Using a weighted k-NN improves significantly the results. There are several types of weights. We mention rang weight (R), linear weight (L) and Dudani’s weight (D). To improve the k-NN performance in another way, we also use the fuzzy k-NN. So, various extensions and improvements of the k-NN rule have been carried out by many researchers. In our work, we focus on Arif *et al.* and Keller *et al.* researchers (Arif *et al.*, 2006). γ and m are parameters of Arif *et al.* (A) and Keller *et al.* (K) rules respectively. We used four different distances: Canberra (C), Discrimination Cost (D), Hamming (H) and Euclidean (E).

In addition, SVM (Mozaffari et al., 2004) can be used with different types of kernel functions $K(x,x_i)$ such as linear (L), polynomial (P) (1), sigmoid (S) (2) and radial basis functions (RBF) (3).

$$K(x, x_i) = (ax.x_i + b)^d \qquad (1)$$

$$K(x, x_i) = \tanh(a(x.x_i) - b) \qquad (2)$$

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \qquad (3)$$

The SVM classifier includes a learning phase (SMO algorithm) followed by a test phase ("one against one" approach). This classifier has three hyper-parameters for these different types:

- C: the Lagrange multipliers maximum;
- T (Tolerance): the prototype class value error;
- E (Epsilon): the Lagrange multipliers error;

3 COMPARATIVE STUDY

If we consider each character form, we attain to more than 120 character classes. But the diacritics elimination reduces the classes’ number to 29. These classes represent the prototypes which can be resulting from segmentation. Diacritics elimination not only reduces the classes’ number but also facilitates the segmentation step, especially for the handwriting when there is an overlapping between diacritics and character body. Example Laaa (لا) are two characters which are Laam (ل) and Alif (ا) and no segmentation technique can segment it, so it is better to recognize it as a class. These classes are presented in Table 1.

Table 1: The different Arabic characters classes.

ـا	ٮـ	ـمـ	ســ	ى
Class 1	Class 2	Class 3	Class 4	Class 5
ـٮ	ں	و	ر	م
Class 6	Class 7	Class 8	Class 9	Class 10
ء	د	ک	ـحـ	ح
Class 11	Class 12	Class 13	Class 14	Class 15
صـ	ط	ه	ـهـ	هـ
Class 16	Class 17	Class 18	Class 19	Class 20
ٯ	ـڡ	ـع	ـعـ	عـ
Class 21	Class 22	Class 23	Class 24	Class 25
ع	لا	ل	س	
Class 26	Class 27	Class 28	Class 29	

In the following, our tests are done on character body (letter). The character recognition can be achieved by taking into account the eliminated diacritics. For the experimental process, we used the half of the data for training and the other for testing randomly. As is shown in previous section, there are many classifiers parameters which are defined experimentally. For fuzzy k-NN parameters, we fixed γ=0.4 and m=0.1. The most difficult step in using SVM is the choice of the appropriate parameters. We tested many parameters and we present those which gave top results. So, we fixed:

- L: $C=10^3$, $T=10^{-2}$, $E=10^{-1}$.
- P: $C=5$, $T=10^{-3}$, $E=10^{-3}$. (a) a=3, b=1, d=2; (b) a=2, b=1, d=3; (c) a=5, b=2, d=4.
- S: $C=10^2$, $T=10^{-3}$, $E=10^{-3}$. (d) a=1, b=3; (e) a=1, b=-4; (f) a=1, b=50.
- RBF: $C=10^3$, $T=10^{-2}$, $E=10^{-1}$. g) σ=0.4; (h) σ=0.6; (i) σ=0.9.

All obtained results are presented in percent (%).

3.1 Printed Dataset

The printed dataset contained 2320 samples (29 classes X 4 fonts X 20 sizes). The fonts were: "Advertising Extra Bold", "Diwani", "Unicode Sara M" and "Times" (see Figure 1). The sizes were: 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44 and 46. This dataset is generated with the same procedure used to generate APTI database (Slimane et al., 2009).

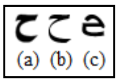

Figure 1: Fonts: (a) Advertising Extra Bold (b) Diwani and (c) M Unicode Sara.

First, we presented performance rates using k-NN. To test distance influence, we fixed k=1. To compare between traditional, weighted and Fuzzy k-NN, we used k=5 and Canberra Distance. To test k variation, we used traditional k-NN and Canberra Distance. We also tested feature combinations. Then, we presented performance rates using SVM. The obtained results are presented in Table 2.

3.2 Multi-oriented Printed Dataset

The multi-oriented dataset contained 9280 samples (29 classes X 2 fonts X 20 sizes X 8 orientations). The fonts were: "Advertising Extra Bold" and "Diwani". We kept the same sizes. The 8 orientations are described in Figure 2.

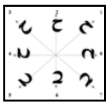

Figure 2: The 8 orientations.

Table 3 presents performance rates using k-NN. The best performance rate 99.26% is given by AMI.

Table 2: Performance rates using printed dataset.

k-NN Classifier							
		ZM	TMM	AMI	FMT	IM	FD
C		99.56	95.77	95.34	93.87	89.65	63.01
D		97.84	95.17	95.77	93.53	88.10	69.74
H		**99.74**	90.60	84.91	94.56	59.91	59.56
E		99.56	89.82	82.67	94.91	56.55	60.94
T		98.53	93.10	90.86	90.68	75.00	47.32
W	R	98.79	94.13	92.32	91.29	81.03	58.53
	L	97.67	94.48	88.87	90.43	74.22	59.65
	D	78.53	95.34	93.36	92.24	82.67	65.34
F	A	99.05	94.39	93.27	92.41	79.31	55.08
	K	98.44	92.84	89.91	88.90	66.46	44.13
k	1	99.56	95.77	95.34	93.87	89.65	63.01
	3	98.96	92.32	94.13	91.37	79.82	54.31
	10	97.50	87.75	89.74	87.67	67.93	33.44
Features Combination + k-NN Classifier (k=1 and Canberra distance)							
		ZM+AMI +FMT	ZM+FMT	ZM+AMI	ZM+AMI+FMT+TMM	ZM+TMM	
Rate		**99.91**	99.82	99.74	99.65	99.31	
SVM Classifier							
		ZM	FMT	FD	IM	AMI	TMM
L		99.56	95.94	83.44	42.24	16.29	3.44
P	(a)	99.48	95.51	84.05	35.60	20.34	3.44
	(b)	99.65	94.91	53.62	30.60	18.96	3.44
	(c)	**99.91**	73.96	3.44	3.44	18.27	3.44
S	(d)	99.56	4.56	2.24	2.80	16.03	2.58
	(e)	99.48	3.53	2.41	2.93	15.94	2.06
	(f)	**99.91**	3.27	3.62	2.50	18.27	2.24
RBF	(g)	99.74	95.60	42.75	40.00	20.25	30.17
	(h)	99.82	95.68	57.32	42.58	16.20	33.27
	(i)	**99.91**	95.43	66.55	44.56	18.01	36.98

Table 3: Performance rates using multi-oriented dataset.

k-NN Classifier (Canberra distance and k=1)						
	AMI	ZM	IM	TMM	FMT	FD
Rate	**99.24**	96.96	93.18	92.52	91.25	62.95
k-NN Classifier						
			AMI	ZM	IM	
k=1	Canberra		99.24	96.96	93.18	
	Discrimination Cost		**99.26**	92.50	92.47	
	Hamming		94.09	97.88	56.20	
	Euclidean		93.31	97.95	49.48	
k=5+ Canberra Distance	Traditional k-NN		96.31	94.67	83.92	
	Weighted k-NN	Rang	97.62	95.84	89.20	
		Linear	98.68	56.48	87.15	
		Dudani	98.85	69.89	89.80	
	Fuzzy k-NN	Arif	98.44	95.60	87.50	
		Keller	99.24	90.30	72.04	
Conberra	k =1		99.24	96.96	93.18	
	k =5		96.31	94.67	83.92	
	k =10		95.53	91.68	77.69	

3.3 Handwritten Dataset

The handwritten dataset contained 2900 samples (29 classes X 100 samples). The samples were chosen randomly from the multi-writers IFN/ENIT dataset (400 different writers). The segmentation and the diacritics elimination were done manually. Figure 3 presents examples from the handwritten dataset.

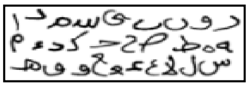

Figure 3: Examples from IFN/ENIT dataset.

Table 4 presents top results using handwritten dataset. The poor performance rates are explained by the dataset nature (handwriting) and the number of writers (400 writers). Despite their weakness, features combination can improve the results by using k-NN (66.68%) as well as SVM (66.62%).

Table 4: Performance rates using handwritten dataset.

<table>
<tr><td colspan="8"><i>k-NN Classifier</i> (<i>Canberra distance and k=1)</i></td></tr>
<tr><td colspan="2"></td><td>ZM</td><td>TMM</td><td>IM</td><td>FMT</td><td>FD</td><td>AMI</td></tr>
<tr><td colspan="2">Rate</td><td><b>45.65</b></td><td>34.13</td><td>32.75</td><td>26.34</td><td>22.55</td><td>22.20</td></tr>
<tr><td colspan="8"><i>Features Combination</i></td></tr>
<tr><td colspan="2"></td><td>All</td><td>ZM+TMM +IM+AMI+ FMT</td><td colspan="2">ZM+TMM +IM+FMT</td><td>ZM+TMM +IM+AMI</td><td>ZM+TM M+IM+F D</td></tr>
<tr><td colspan="2">Rate</td><td><b>66.68</b></td><td>66.13</td><td colspan="2">65.10</td><td>62.06</td><td>61.10</td></tr>
<tr><td colspan="8"><i>SVM Classifier</i></td></tr>
<tr><td colspan="2"></td><td>ZM</td><td>FMT</td><td>FD</td><td>AMI</td><td>IM</td><td>TMM</td></tr>
<tr><td colspan="2">L</td><td>55.24</td><td>33.44</td><td>17.03</td><td>14.27</td><td>13.1</td><td>3.44</td></tr>
<tr><td rowspan="3">P</td><td>(a)</td><td>60.34</td><td>26.62</td><td>20.55</td><td>11.17</td><td>9.10</td><td>3.44</td></tr>
<tr><td>(b)</td><td>60.62</td><td>30.96</td><td>17.79</td><td>11.86</td><td>6.41</td><td>3.44</td></tr>
<tr><td>(c)</td><td>57.72</td><td>37.37</td><td>3.44</td><td>14.82</td><td>3.44</td><td>3.44</td></tr>
<tr><td rowspan="3">S</td><td>(d)</td><td>50.89</td><td>3.86</td><td>2.96</td><td>7.37</td><td>3.51</td><td>3.65</td></tr>
<tr><td>(e)</td><td>54.20</td><td>3.31</td><td>3.44</td><td>8.27</td><td>3.70</td><td>3.65</td></tr>
<tr><td>(f)</td><td>51.93</td><td>3.44</td><td>3.79</td><td>7.58</td><td>3.03</td><td>3.51</td></tr>
<tr><td rowspan="3">R
B
F</td><td>(g)</td><td>60.68</td><td>37.17</td><td>18.62</td><td>16.20</td><td>7.86</td><td>4.68</td></tr>
<tr><td>(h)</td><td><b>60.89</b></td><td>38.13</td><td>23.03</td><td>14.13</td><td>7.72</td><td>4.68</td></tr>
<tr><td>(i)</td><td>60.34</td><td>30.48</td><td>22.89</td><td>12.68</td><td>8.06</td><td>4.75</td></tr>
<tr><td colspan="8"><i>Features Combination</i></td></tr>
<tr><td colspan="2"></td><td>ZM+FM T</td><td>ZM+FMT+ FD</td><td colspan="2">ZM+FMT+ FD+AMI</td><td>ZM+FD</td><td>All</td></tr>
<tr><td colspan="2">L</td><td>61.86</td><td>61.17</td><td colspan="2">62.13</td><td>54.96</td><td>3.44</td></tr>
<tr><td colspan="2">(h)</td><td><b>66.62</b></td><td>37.44</td><td colspan="2">35.51</td><td>32.62</td><td>5.03</td></tr>
<tr><td colspan="2">(b)</td><td>26.82</td><td>39.44</td><td colspan="2">40.62</td><td>38.20</td><td>3.44</td></tr>
</table>

4 CONCLUSIONS AND FUTURE WORK

The present paper proposed a comparative study over Arabic optical character recognition, following the statistic approach. We tried to highlight the obtained results using different datasets, different feature extraction techniques and different classifiers. For printed and for handwritten datasets, Zernike moments give the best recognition rate. This conclusion can be explained by Zernike polar coordinates which are more robust than other coordinates types. For multi-oriented dataset, affine moment invariants are in first position. This conclusion can be explained by their robust invariance to the rotation. The choice of k-NN or SVM depends on the system needs. In future experiments, we aim to extend our study to larger datasets and to incorporate and to study other different feature extraction techniques (wavelets, fractal dimension ...) and different classifiers (neural networks...). In future work, we will develop the system towards Arabic words and texts recognition.

REFERENCES

Abandah, G. and Anssari, N., 2009. Novel moment features extraction for recognizing handwritten arabic letters. *J. Comput. Sci., 5: 226-232.* DOI: 10.3844/ jcssp. 2009. 226.232.

Abdul Sattar, S., Shah, S., 2012. Character Recognition of Arabic Script Languages. *ICCIT*, pp. 502-506.

Aboaisha, Hosain, Xu, Zhijie and El-Feghi, Idris (2012) An investigation on efficient feature extraction approaches for Arabic letter recognition. In: PQDJCEAR' Conference 2012: CEARC'12, pp. 80-85. ISBN 978-1-86218-106-9.

Arif, M., Brouard, T., Vincent, N., 2006. A new fuzzy k-Nearest Neighbors rule. Pattern recognition.

Imran, K. P., Abdulbari A. A., Ramteke R. J., 2012. Recognition of Offline Handwritten Isolated Urdu Character. ACR ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 4, Issue 1, pp.-117-121.

Jenabzadeh, M. R., Azmi, R., Pishgoo, B., Shirazi, S.S., 2011. Two Methods for Recognition of Hand Written Farsi Characters. *IJIP, Vol. 5*(4).

Mozaffari, S., Faez, K., Kanan, H., R., 2004. Feature Comparison between Fractal Codes and Wavelet Transform in Handwritten Alphanumeric Recognition Using SVM Classifier. *ICPR (2)*: 331-334.

Sabri A. Mahmoud, Ashraf S. Mahmoud, 2009. Arabic Character Recognition Using Modified Fourier Spectrum (MFS) vs. Fourier Descriptors. Cybernetics and Systems 40(3): 189-210.

Slimane, F., Ingold, R., Kanoun, S., Alimi, M. A., and Hennebert, J., 2009. A New Arabic Printed Text Image Database and Evaluation Protocols. *In proc. of 10th IEEE ICDAR 2009*, *Barcelona* (Spain), July 26 - 29 2009 , pp. 946-950.

Zaghloul, R. I., AlRawashdeh, E., F., Bader, D. M. K., 2011. Multilevel Classifier in Recognition of Handwritten Arabic Characters. *Journal of Computer Science 7* (4): 512-518.