Arabic Language Resources and Evaluation – Status and Prospects, LREC2002

1st June 2002. Las Palmas de Gran Canaria.

Arabic Character Recognition using Approximate Stroke Sequence

Professor Mohammed Zeki Khedher* **Dr. Gheith Abandah****

*Dept. of Electrical Engg khedher@ju.edu.jo **Dept. of Computer Engg abandah@ju.edu.jo

Jordan University, Amman - Jordan

1. Introduction

Automatic recognition of handwriting has become a mature discipline at the beginning of the 21st century. On-line systems are now available for handheld computers with an acceptable performance. Off-line systems are less accurate than on-line systems. However they are now good enough for specialized systems such as interpreting handwritten postal addresses on envelopes and reading courtesy amounts on bank checks(Plamondon 2000).

The recognition of Arabic characters is particularly difficult due to the necessity of segmentation even for printed text. An Arabic character has up to four different shapes; the shape of a character depends on the type of characters to its left and right.

2. Main characteristics of Arabic writing

Arabic text is written from right to left and is always cursive. Arabic characters are used with some modification in many languages other than Arabic, e.g. Urdu and Persian. The shape of the Arabic character changes according to its location in the word. The Arabic character set is composed of 28 basic characters. 15 of them have dots and 13 without dots. Dots above and below the characters play a major role in the distinction between some characters as they differ only by the number of dots or their locations. Take the example of the letters : ب ت ث ي ن . All these five letters are written in their middle form in the same way as: بـ تـ ثـ يـ نـ . They differ only by the number of dots and their locations.

There are 3 more characters which augment the basic Arabic characters. Those are the "Hamza ء " , "Taa marbuta ة " and "Alif maqsura ى".

There are four characters which may have an association with a the secondary character "Hamzah ء". Those are "Alif أ إ ", "Waw ؤ " , "Yaa ئ " and "Kaf ك ".

There are some more secondary characters used above and below the characters to indicate vowels but we shall exclude them now from our discussions.

Arabic characters do not have fixed width or fixed size, even in printed form.

2.1 An important phenomena in Arabic writing

Arabic writing is known to be cursive even in printed form. However it differs from cursive handwriting of English or other languages of Latin origin by that some characters can be connected from one side only. Out of the 28 basic Arabic characters,

six can be connected from the right side only while the other 22 can be connected from either left or right directions. These six characters are : dal (د), raa (ر), waw (و), alef (ا), thal (ذ) , and zay (ز). These six characters have only forms , the stand-alone form and the final form.; while the rest of the characters have can appear in any of four forms, the initial, the middle, the final and the stand-alone form.This imposed upon the Arabic word the phenomena that it may consist of one or more sub-words. A sub-word can be defined as the basic stand-alone pictorial block of the Arabic writing. Any optical character processing of Arabic characters should treat the sub-word as the basic block for processing whatever the method it uses for preprocessing, segmentation, recognition or classification. This is because each sub-word is separated from other sub-word by a space. Although spaces between sub-words is usually shorter than that between successive words, still it is surrounded by a space. A word may contain one or more sub-words. Some of these sub-words may even consist of a single character in its stand-alone form. Hence, their recognition does not need segmentation.

Shape of the letter in the text differs according to the location of the character in the sub-word i.e. a character at the end of sub-word has exactly the same shape when it comes at the end of a full word.

2.2 Test Sample an Results

In order to give a fair idea about the importance of this phenomena, A sample of Arabic text consisting of about 750kb was chosen. It was randomly selected. Statistics presented here about this sample text, may give a rough picture about the structure of Arabic words in terms of sub-words and the four types of characters.

Table (1)

ers Per Sub-work	No. of	Sub-words%	No. of stand-alone characters	No. of initial characters	No. of middle characters	No Ch
1	147060	46.5%	147060	0	0	
2	91992	29.1%	0	91992	0	
3	47462	15%	0	47462	47462	4
4	22490	7.1%	0	22490	44980	2
5	5949	1.9%	0	5949	17847	
6	1120	0.3%	0	1120	4480	
7	164	0.05%	0	164	820	
8	13	0.004%	0	13	78	

of characters	316250	100%	147060	169190	115664	1
of characters			24.5%	28.1%	19.24%	2

Table (1) shows the analysis for this sample.
It is shown that the sample consisted of 147785 words with 601107 characters. This means that the average word length is 4.1 character per word. The number of sub-words is 316250. This means that on the average there is 1.9 sub-word per word. In other words on the average each word contains about 2.1 sub-words.The number of sub-words consisting of one character is 147785 which makes about 46.5% of the total number of sub-words (and about 25% of the total number of characters). This means that, in the process of optical character recognition, slightly less than one half of the sub-words need no segmentation at all (whether printed or handwritten). Number of sub-words which consists of two characters is 91992 which makes 29.1% of the total number of sub-words. This means that about 30% of the total number of sub-words need segmentation in two characters only. The table also shows that on the average the four different shapes of characters are almost equal with the middle form slightly less (24.5% stand-alone, 28.1% for each of the initial form and the final form, as the number should be equal, and 19.24% for the middle form).

2.3 Proposal for a new procedure for recognition of Arabic characters
According to the above discussion, the approach suggested therefore here is to separate the text into three groups:

1. Sub-words consisting of one character that is in the stand-alone form. This is to be recognized directly without necessity of segmentation.
2. Sub-words consisting of two characters. The first one is in the initial form and the second one in the final form. This needs segmentation in two parts only. If there is a pre-knowledge of the number of characters to be segmented in the sub-word, then the task should be easier.
3. Sub-words consisting of more than two characters. The first one is in the initial form, the last one in the final form, and the rest are in the middle form

Figure 1, shows the flow diagram for this procedure

Segmentation of text into words

Segmentation of words into sub-words

There or more ←---------- No. of characters per sub-word -------→ One

Two

Segment the sub-word into first character in the initial form, the last in the final form and the rest of the characters(1 or more) in middle form	Segment the sub-word into two characters, the first one in the initial form and the the second in the final form	Recognize the characters with-out segmentation as stand-alone form

Figure 1
Proposed procedure for Arabic Character Recognition

3. previous Work in Arabic Character Recognition

Several good literature review papers were published for various research on Arabic character recognition(Jambi, 1991,Ahmed, 1994, Amin, 1997 and Ahmed, 2000]
Here we shall give some elaboration on some of the effort spent in this direction.
Classical optical off-line recognition of handwriting is composed of a preprocessing stage, character segmentation, feature extraction and a classification stage(Casey 1996). Preprocessing consists of several operations, like, thresholding, noise removal, page orientation and skewing of lines removal, line segmentation, word segmentation, and pictures and figures removal,

Preprocessing of character is an important part of the OCR systems whether it is for printed text or handwritten. Thinning of the script is an important step in many systems. There is no specific characteristic of Arabic writing to make such process different. Detection of tree, end-point and edge-points was considered as an important step in preprocessing of handwritten Arabic recognition(Abdulla, 1988). Clustering techniques for skeltonization was addressed by some authors(Mahmoud, 1991). Normalization and orientation of Arabic script were the topic of discussion of some literatures(Hissain & Cowel, 2000). Geometrical modeling of some Arabic characters were also tried(Hussain & Zalik, 2000).
Work on isolated printed Arabic characters which take into account the overall location of the character compared with the baseline. Some authors(Talba 1987) tried to study the location of different characters, e.g. whether located in a middle region, in the middle and upper region; in the middle and the lower region or the upper, middle and lower locations at the same time.

The chain code describing the sequence of character strokes using the 8-directions strokes was followed by the majority of researchers. However hexagonal sampled procedure was also found(Khellah, 1994).
Selection of proper features may helps in better recognition. Features of round shapes, found in Arabic handwriting, are some of these(Badie, …).
A 1-D slices of printed character spectrum is presented and the estimation of the projection of the X- and Y- axes is estimated. Features were extracted from the 2-D spectrum and compared with model's features using a distance measure. Results showed good recognition rates(Alshebeli, 1997).
Segmentation of characters is an important step in character recognition for cursive writing whether handwritten of printed. There are three strategies for segmentation. They are: the classical approach in which segments identification is based on "character-like" properties, the recognition-based segmentation strategy, in which the system searches the image for components that match classes in its alphabets, and the third strategy is the holistic method, in which the system seeks to recognize words as a whole, thus avoiding the need to segment into characters(Casey 1996).

Segmentation of Arabic printed characters was treated in many ways. The hierarchical syntactic procedure is one of these approaches(Haj-Hassan, 1990). Some others used statistical approach and moments of horizontal and vertical projections were used to enable classification using quadratic discriminating functions(Udpa, 1992). It was tried by tackling both the horizontal and the vertical overlapping. Horizontal and vertical histograms were used to get primitives of the sub-words(Abo Samra, 1997). Even segmentation of printed Arabic characters was tried without the thinning process. This was achieved by tracing of the outer contour of the sub-word and calculating the distance between the extreme points of intersection of the contour with the vertical line(El-Sheikh, 1988). Separation of Arabic printed characters have been performed using the moment invariant algorithm, and found good rate of recognition(El-Khaly, 1990).

Mathematical correlation between the unknown character features and those of the standard characters proved to have some indication towards classification of characters into groups which may be recognized at a subsequent stage(Nouh, 1988).

Recognition of handwritten cursive Arabic characters where dealt with in many literature. Structural features where selected and a suitable clustering technique was chosen for classification(El-Desouky, 1992). Others used a tree representation for the handwritten characters for the description of various characters(Saleh, 1994 & Saleh, 1996). Use of tree representation and fuzzy constrained graph models which tolerate large varieties in writing styles were reported also(Abuharba, 1994).

The recognition of Multi-font Arabic writing using neural network has been dealt with in some publications(Altuwaijri, 1995).
A comprehensive review of Arabic Character Recognition techniques and evaluation of status of such systems development is given(Ahmed, 2000)
Recognition of different fonts of Arabic printed text was tried using preprocessing and structural feature extraction. Results showed good recognition rates. However the recognition of dots showed low recognition rated(Kavianifar,1998).

Full speed and high accuracy were the goal for proposal of parallel Arabic OCR systems. There were no concrete results to prove the feasibility of such systems(Alherbish, 1997).

Hidden Markov Models which proved to be very successful in the are of automatic speech recognition was tried in the area of omnifont open-vocabulary Arabic OCR system. Character error rates around 1% was reported(Bazzi, 1999).

Accumulative invariant moment was used as an identifier in character recognition(El-Dabi, 1990). However the high sensitivity of the procedure, does not enable it to be applied for handwriting character recognition.

Context-free and tree grammars methods were tried on several handwritten Arabic characters and both showed promising results(Al-Waily, 1989).

Work on limited handwritten Arabic text database was tried. A system based on four types of basic features, namely the end points, corners, the strokes and the branch points gave reasonable results(Jambi, 1991).

Recognition of Typewritten Arabic characters gave good results using external features such as character area ratio, n-th quadrant ratio, vertical line ratio, horizontal line ratio, number of upper edges and other similar features(Al-Ohali, 1995).

Online character recognition uses the feature extraction process results in a sequential manner which is called the chain code. Treatment of secondary characters (mainly the points above and below the characters) is definitely an integral part of the recognition process(El-Gwad, 1990).

Neural Network was used for the recognition of handwritten Arabic Numerals. Results were promising(Said, 1998 , Al-Kadi, 1995 and Al-Sharaidah, 2000).

4.Approximate Stroke Sequence String Matching

Given two character images, there is no universally accepted definition for similarities or differences between them . If the two images are converted into a one-dimensional

string, then the task will be easier to define. Distance between two histograms of angular measurements was the subject of some literatures(Cha 2000).

The stroke sequence is based on the 8 types of strokes so that → takes the value of 0 and the 45 degree direction takes the value of 1 and ↑ takes the value of 2 ,← the value of 4 , ↓ the value of 6 and the values of 3,5 and 7 are in between.
Stroke sequence string matching(Cha 1999) is based on the individual distance $d_{i,j}$ between the i'th stroke in letter a_1 and the j'th stroke in letter a_2 where

$d_{i,j} = |a_1(i) - a_2(j)|$ if $|a_1(i) - a_2(j)| \leq 4$

This value is to be modified so that $d_{i,j}$ value is replaced by the value $8-d_{i,j}$ when it exceeds the value of 4. This is as taking the smallest distance between the two directions whether the rotation is clockwise or counterclockwise. Hence $d_{i,j}$ gives the minimum number of necessary steps to turn from the direction given by $a_1(i)$ and $a_2(j)$.
Allowing a cost function of c=2 which allows a minimum edit distance between two stroke sequence strings.
Figure 2a and 2b show the Arabic characters ع and ح taken as standard shapes. The stroke sequences corresponding to each of them is shown in the upper row of Tables 2 and 3 respectively. Figure 2c shows the same for a string supposed to be of an unknown character to be compared to each of them. . However this is a handwritten letter ع. The stroke sequence for this character is shown in the left hand column of both Table 2 and 3 Tables 2 and 3 show the computing distance tables between the unknown character and each of the two characters ع and ح. It is shown that the distance between the unknown character and the ع is 8 , while its distance to the ح is 10. This gives the result that the unknown character is ع and not ح .
The individual strokes on the top and left hand side of the Table (starting row and column) are t_i and l_j simultaneously. The calculations involved in the shown Tables takes into consideration that
T(i,j) is the minimum value of T(i-1 , j-1) $+d_{i,j}$ and the two values T(i-1, j) +c when t_i is missing and T(I, j-1) +c when l_j is missing.

Table 2

String matching between unknown character and character ع

		2	0	0	5	5	6	7	0	1	2
	0	2	4	6	8	10	12	14	16	18	20
2	2	0	2	4	6	8	10	12	14	16	18

4	4	2	4	6	5	7	9	11	13	15	17
5	6	4	5	7	6	5	7	9	11	13	15
6	8	6	6	7	8	7	5	7	9	11	13
0	10	8	6	6	8	9	7	6	7	9	11
4	12	10	8	8	7	9	9	8	9	10	11
5	14	12	10	10	8	7	9	10	11	12	13
6	16	14	12	12	10	9	7	9	11	13	15
7	18	16	14	14	12	11	9	7	10	12	14
6	20	18	16	14	14	13	11	9	7	9	11
2	22	20	18	16	16	15	13	11	9	8	9
4	24	22	20	18	17	17	15	13	11	10	10

Table 3

String matching between unknown character and character ح

		3	4	5	6	7	0	4	5	5	6	7	0	1
	0	2	4	6	8	10	12	14	16	18	20	22	24	26
2	2	1	3	5	7	9	11	13	15	17	19	21	23	25
4	4	3	1	3	5	8	9	11	13	15	17	19	21	23
5	6	5	3	1	3	5	7	9	11	13	15	17	19	24
6	8	7	7	3	1	3	5	7	9	11	13	15	7	19
0	10	9	9	5	3	2	3	5	7	9	11	13	15	17
4	12	11	9	7	5	4	5	3	5	7	9	11	13	15
5	14	13	11	9	7	6	7	5	3	5	7	9	11	13
6	16	15	13	11	9	8	8	7	5	4	5	7	9	11
7	18	17	15	13	11	9	9	9	7	6	5	5	7	9
0	20	19	17	15	13	12	11	11	9	8	7	6	4	7
2	22	21	19	17	15	14	13	13	11	10	9	8	7	6
4	24	23	21	19	17	16	15	13	13	12	11	10	9	8

5. Results and Discussion

The above algorithm was applied to various Arabic stand-alone characters collected from 48 different persons. Figure 3 shows such character from the different handwriting of the 48 persons for the letter ع. These persons were instructed to copy the same text without any restrictions on their writing style. The handwritten pages were scanned and then normalized to nearly the same overall sizes. The stand-alone characters were then copied and analyzed by the above algorithms. Results showed recognition rate of about 80% for characters such as ا , ص , ه while the rate of recognition was much less than that for nearby characters such as ب & ف or ق & ن .

6. Conclusions

A statistical analysis on a sample of Arabic text, showed that average Arabic word contains about 4 characters in an average of 2 sub-words. This shows that the basic block to be dealt with in Arabic OCR systems should be the sub-word rather than the word. The size of the sub-word varies from a single character usually up to 8 characters. The method of recognition of sub-words of different length ought to be different. For sub-words with a single character, the matching for recognition may be made only with characters of the stand-alone form. For sub-words with two characters only, a single shot segmentation has to be made dividing the sub-word into two characters. The first one is of the initial form and the second one of the final form. Sub-words of length longer than 2 characters, need to be segmented into three characters or more. The first is of initial form, the last of final form and the rest of middle form. Design of Arabic OCR system when taking these facts into account shall be much simpler. However the classification of sub-words according to the number of the characters they contain, still ought to be addressed.

As a first step, the recognition of the sub-words with a single character of a stand-alone form has been treated using the approximate stroke sequence string matching. Promising results are shown. Further refinement of the algorithm use need to be carried out for better rate of recognition.

7. References

Abd El-Gawad A.O., M.M. Salem, F.E.Z. Abou Shadi and H. Arafat, 1990. Automatic Recognition of Handwritten Arabic Characters. *10th International Conference on Pattern Recognition*.

Abdulla W.H., A.O.M. Saleh and A.H.Morad, 1988. A Preprocessing Algorithm for Hand-Written Character Recognition. *Pattern Recognition Letters.* 7 , 13-18.

Abo Samra G., K. Jambi, H. Al-Barhamtoshy, R. Amer and I. Al-Bidewi, 1997. A Comprehensive Algorithm for Segmenting Handwritten Arabic Script in Off-line Systems. *17th NCC, KFUPM.* 1-13.

Abuharba I.S.I., S.A. Mahmoud and R.J. Green, 1994. Recognition of Handwritten Cursive Arabic Characters. *HLL Transaction of Pattern Analysis and Machine Intelligence.*16:6. 664-672.

Ahmed P. and Y. Al-Ohali, 2000. Arabic Character Recognition: Progress and Challenges. *Journal of King Saud University.* 12, Comp. & Inf. Sci. 85-116

Ahmed P. and M. A. A. Khan, 1994. Computer Recognition of Arabic Script Based Text: the State of the Art. *Proc. of the 4th International Conference and Exhibition on Multi-lingual Computing (ICEMCO 94, Arabic and Roman Script)*, London, 22.1-22.15..

Al-Ohali Y.A.M, 1995. *Development and Evaluation Environment for Typewritten Arabic Character Recognition*. M.Sc. Thesis, King Saud University. Riyadh.

Al-Sharaidah, M. A., 2000. *Recognition of Handwritten Arabic Characters via Neural Networks* M.Sc. Thesis, Jordan University.

Al-Yousefi H. and S.S. Udpa, 1992. Recognition of Arabic Characters*," IEEE Transaction on Pattern Analysis and Machine Intelligence.* 14: 8. 853-857.

Altuwijiri M.M. and M.A. Bayoumi, 1995. A New Recognition System for Multi-Font Arabic Cursive words. *Proceedings of ICECS'95*, Amman, Jordan. 298-303.

Amin A., 1997. *Handbook of Character Recognition and Document Image Analysis," Chapter 15* World Scientific Pub. Co. 397-420.

Alshebeili S.A., A.A.F Nabawi and S.A. Mahmoud, 1997. Arabic Character Recognition Using 1-D Slices of the Character Spectrum. *Signal Processing, Elsevier.* 56. 59-75.

Alherbish J. and R.A. Ammar, 1997. Arabic Character Recognition in a Multi-processing Environment. *Proceedings of the 2nd IEEE Symposium on Computers and Communications.*

Al-Kadi Z. and S. Serhan, 1994. The Use of Learning Networks in Recogniton of Arabic Numeral. *Dirasat*, Jordan University. 22B:4. 933-949.

Al-Waily R.S.A., 1989. *A Study on Preprocessing and Syntactic Recognition of Hand-Written Arabic Characters*. M.Sc. Thesis, Basrah University, Iraq.

Plamondon, R., 2000. On-Line and Off-line Handwriting Recognition, A Comprehensive Survey. *IEEE Transaction on Pattern Analysis And Machine Intelligence,* 22:1.

Casey R.G. and E. Lecolinet, 1996. A Survey of Methods and Strategies in Character Segmentation. *IEEE Transaction on Pattern Analysis And Machine Intelligence.* 18:7.

Cha S.H., Y.C. Shin and S.N. Srihari, 1999. Approximate Stroke Sequence String Matching Algorithm for Character Recognition and Analysis. *5th International Conference on Document Analysis & Recognition* . Banglore, India

Cha S.H. and S.N. Srihari, 2000 "Distance between Histograms of Angular Measuremnts and its Application to Handwritten Character Similarity," *Proceedings of International Conference on Pattern Recognition.*

Talba M.F., S.A. Wahba and A. Salim, 1987. A Recognition Algorithm for Arabic Printed Characters," *Proceedings of the International Symposium of Applied Informatics,* Switzerland. 128-131

Nouh A., A.N. Ula and A. Sharaf Eldin, 1988 Algorithm for Feature Extraction: A case Study for the Arabic Character Recognition. *Proceedings of 10th National Computer Conference,* Saudi Arabia. 653-666.

Bazzi I., R. Schwartz and J. Makhoul, 1999. An Ominfont Open-Vocabulary OCR System for English and Arabic. *IEEE Transaction on Pattern Analysis and Machine Intelligence.* 21: 6. 495-504.

Badle K. and M. Shimura, Machine Recognition of Arabic Cursive Scripts. *Pattern Recognition in Practice* . North-Holland Pub.
El-Khaly F. and M.A. Sid-Ahmed, 1990. Machine Recognition of Optically Captured Machine Printed Arabic Text. *Pattern Recognition.* 23: 11, 1207-1214.
Sami S.E., R. Ramsis and A., 1990. Arabic Character Recognition System, A Statistical Approach for Recognizing Cursive Typewritten Text. *Pattern Recognition*. 23: 5, 485-495.
Haj-Hassan F., 1990. Arabic Character Recognition. *Computers and the Arabic Language*, Hemishere Publishing Co. 113-118.
Jambi K. M., 1991. Arabic Character Recognition: Many Approaches and One Decade. *The Arabian Journal for Science and Engineering.* 16: 4B. 499-508.
Mahmoud S. A., I. AbuHaiba and R.J. Green, 1991 Skeltonization of Arabic Characters Using Clustering Based Skeltonization Algorithm. *Pattern Recognition.* 24: 5. 453-464.
El-Desouky A.I., M.M. Salem A.O. Abd El-Gwad and H.Arafat, 1992. A Handwritten Arabic Character Recognition Technique for Machine Reader. *International Journal of Mini and Microcomputers.* 14:2. 57-61.

Salih A.O.M. and R.S. Al-Waily, 1994. Tree Representation of Hand-writen Arabic Characters. *Mu'tah Journal for Research and Studies.* 9:3. 125-140.

Khellah F., 1994. Recognition of Hexagonally Sampled Printed Arabic Characters. *The Arabian Journal for Science and Engineering.* 19:4A. 565-585.

Salih A.O.M. and R.S. Al-Waily. 1996. Recognition of Hand-Writen Arabic Characters sing Tree Grammars *2nd Jordanian Conference in Electrical Engineering, Mutah University.* 402-413.
Kavianifar M. and A. Amin, 1998. Preprocessing and Structural Feature Extraction for a Multi-Font Arabic/Persian OCR. *Proceedings of 5th International Conference of Document Analysis and Recognition.*

Said F. N., R.A. Yacoub and C.Y. Suen, 1998. Recognition of English and Arabic Numerals Using a Dynamic Number of Hidden Neurons. *Proceedings of 5th International Conference of Document Analysis and Recognition.*

Hussain F., B. Zalik and S. Kolmanic, 2000. Intelligent Digitisation of Arabic Characters. *Proceedings of the Conference on Information Visualization.* 337-342.

Hussain F. and J. Cowell, 2000. Character Recognition of Arabic and Latin Scripts. *Proceedings of the Conference on Information Visualization.* 51-56.

El-Sheikh T.S. and R.M. Guindi, 1988. Computer Recognition of Arabic Cursive Script. *Pattern Recogniton.* 21:4. 293-302.

Jambi K.M., 1991. *Design and Implementation of a System for Recognizing Arabic Handwritten Words with Learning Ability*. Ph.D. Thesis , Illinois Institute of Technology.

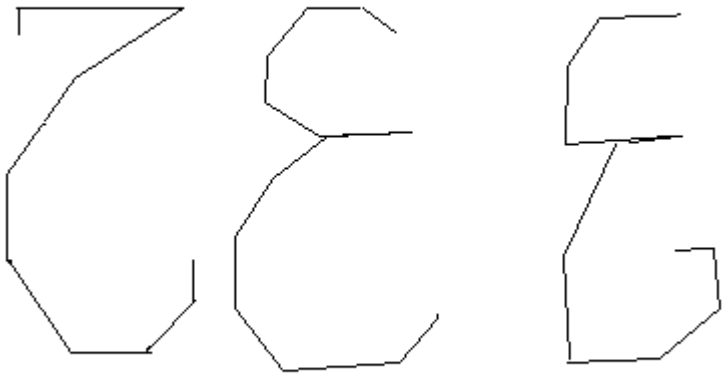

Figure 3 The writing of 48 different persons of the same character ع