Информационные процессы, Том 6, № 1, 2006, стр. 11–19

Preprocessing phase for Arabic Word Handwritten Recognition

Hasan Al-Rashaideh

Saint Petersburg Institute for Informatics and Automation, Russian Academy of Sciences,
14 Line, 39,199178, Saint Petersburg, Russian Federation
http://spbrc.nw.ru/english/main/index.htlm

Received February 26, 2006

Abstract—In this paper we reviewed the importance of the pattern classification and its application. We list the characteristics of Arabic language writing style, furthermore focused on the preprocessing step of the recognition system. We described and tested algorithm to create skeleton which will be the base representation of Arabic words which we will use for feature extraction phase. Also we discussed and implemented the algorithm of baseline detection. The algorithms of skeleton and baseline detection are tested using database IFN/ENIT of handwritten Tunisian town names and they work properly.

1. INTRODUCTION

The development of the computer technologies is become more flexible and applicable in a wide range of areas, furthermore automation of the Companies' works now are placed and seen everywhere around us.

Problem of transformation scanned documents which contain handwritten texts to computer for further processing still needed for automation office's tasks in big companies, for form processing applications mainly in bank check reader systems and post office address reader systems.

Achievements in technologies for automated systems orientated to English-like languages nowadays are growth and work well in real life, while it's still active research in other languages such as Arabic. Beside that it's difficult to apply available algorithms for English directly to other languages.

Here we focused on the Arabic writing style and its characteristics and developed applicable algorithms for the specifics of the handwriting in Arabic.

2. GENERAL BACKGROUND

Pattern classification is a scientific discipline for categorize data into distinguishable classes. Many sciences required classification processes such as Computer vision, Artificial intelligence, Data Mining, Multimedia Information Retrieval, Medicine, and Biology.

Measurements and observations are selected from input data to form the pattern through feature extracting process. Then these features are used for classification process, in addition this operation is the core of any recognition system and must be analyzed and designed carefully to get high rate of recognition.

Any recognition system must have two main stages:

– Feature extraction: extract measurements from the input to distinguish between classes, Devijver and Kittler in [1] suggested that the problem of extracting Features from input data is done by

selecting information which is most relevant for classification purposes and able to discriminate between classes, in the sense of minimizing the within-class pattern variability and maximize the between classes pattern variability.

– Classification process: Determine the class to which the input belongs; figure 1 illustrates pattern recognition system, and this stage done through one of the following approaches: Statistical pattern recognition and syntactic pattern recognition, and neural pattern recognition.

The goal of handwritten recognition is to identify an input character's image correctly; and it's an application area of Pattern Recognition, in addition the conversion of handwritten text image to editable form is important to many automated processing systems. Two main techniques

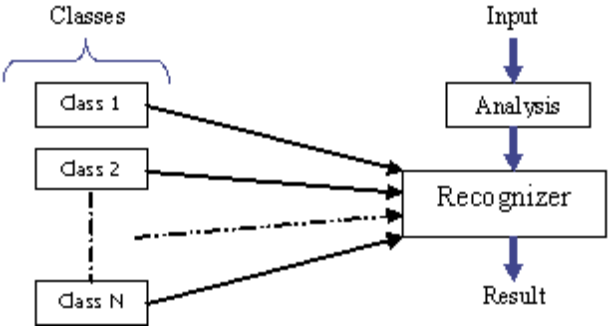


Figure 1. Overview of Pattern Recognition System

are used for character recognition depending on the intended application. The first is off-line or optical character recognition (OCR). The system accepts its input from a digital scanner or from a picture using some image processing algorithm needed before classification step. The alternative is online character recognition; here the system accepts input data in real time and then computes the relationships between points to extract the features in real time.

Brief comparison between offline and online approaches:

– Online recognition system: the system accepts the movement of pen from the hardware such as graphic tablet, light pen; and there is a lot of information during the input process available such as: current position, movement's direction, stopping points, starting points, strokes order. Al-Taani In [2] proposed online Arabic digital recognition system.
– Offline recognition system: the system accept image as input from scanner, offline recognition in more difficult than online character recognition: because of not availability of contextual information and prior knowledge like text position, size of text, order of strokes, stop points, and start points, furthermore there are a noises in image while the noises in online recognition near to be absent. A comprehensive papers [3, 4, 5] disscused the scientific progress in the Offline word recognition.

3. ARABIC CHARACTERS

Arabic characters are used in writing many languages not only in Arabic countries, but for Urdu and Farsi and other languages in countries where Islam is the principal religion (e.g., Iran, Pakistan, and Malaysia). The special characteristics of Arabic written words and characters do not allow the direct application of algorithms for other languages. See figure 2.

خ	ح	ج	ث	ت	ب	أ
Xaa'	*H'aa'*	*Jeem*	*Thaa'*	*Taa'*	*Baa'*	*'Alif*
ص	ش	س	ز	ر	ذ	د
Saad	*Sheen*	*Seen*	*Zaay*	*Raa'*	*Thaal*	*Daal*
ق	ف	غ	ع	ظ	ط	ض
Qaaf	*Faa'*	*Ghayn*	*'Ayn*	*Thaa'*	*Taa'*	*Daad*
ي	و	هـ	ن	م	ل	ك
Yaa'	*Waw*	*Haa'*	*Noon*	*Meem*	*Laam*	*Kaaf*

Figure 2. Stand-alone Arabic Characters

Arabic's Letters characteristics are, see figure 3 :

- Arabic is a cursive type language written from right to left.
- Arabic has 28 basic characters. Each character has 2-4 forms depending on its position within the word.
- Many letters of the Arabic alphabet have dots , above or below the character body , and some letters have a Hamza (zigzag shape)and Madda.
- Overlapping characters: some Arabic's characters become over each other horizontally when they connected with each other.

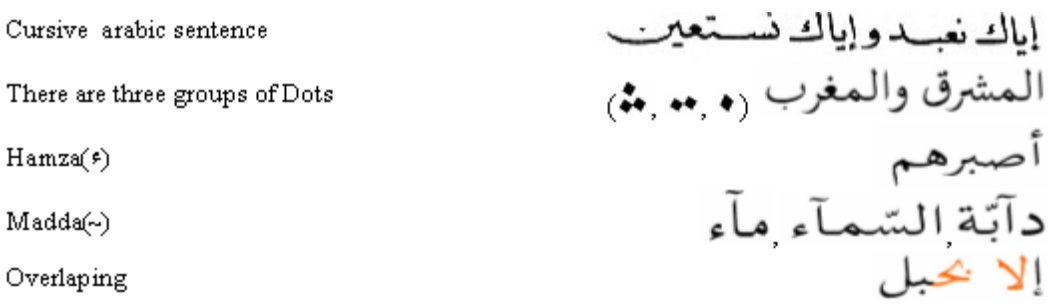


Figure 3. Some charactaristics of Arabic characters with examples

4. RECOGNITION SYSTEM

Any Handwritten recognition system typically consists of the following phases [4]: scanning, digitization, preprocessing, segmentation to single character or segments related to character, feature extraction, recognition using classifier, post processing for verification using lexicon and the last step of the system is evaluation. in figure 4 shows a typical offline character recognition system and it's components. Digitizing consist: noise elimination, gaps filling, size translation, normalization, and binarization. While preprocessing phase convert the original form of input *image* to another form *skeleton* to simplify feature extraction phase.

To build an application for handwritten text, all phases of the automated handwritten recognition system should be designd carefully and precisely because of the varibility and complxity of the

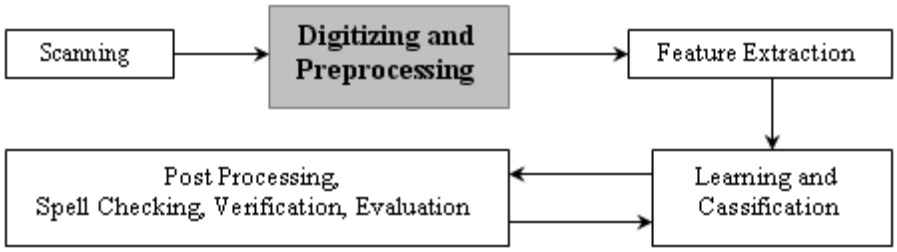


Figure 4. Offline Character Recognition System

problem. One of the most important stages is preprocessing phase, which constructs the representation of the character and word in order to simplify the feature extraction phase, in this research we concentrated on two helpful substages: Skeleton finding through thinning, Baseline identification.

4.1. Dataset

For experiments purposes we used database IFN/ENIT of handwritten Tunisian town names which collected by Pechwitz [6], Database contain 946 town names and written by 411 writers. Figure 5 shows an example from the dataset. Words' images in database are already binary images, and also noise elimination, gap-filling and normalization all are done.

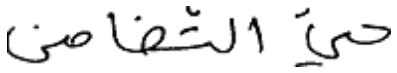


Figure 5. Town name example from IFN/ENIT Dataset

4.2. Thinning

the process to reduce image size to compact size and find the medial axis which defines as a set of pixels S where these pixels have an equal distance from the boundary pixels around it, and the output of this process is skeleton for the handwritten word, this process must save the geometry and the connections between the characters and the location of original character[7, 8], based on border pixels removing recursively taking into account saving the geometry, location and connections.

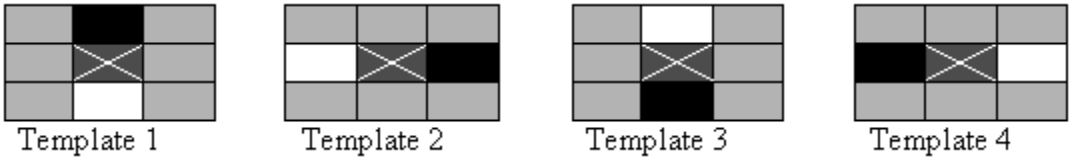
Skeleton representation advantages:

- Good way to represent the structural relations between components in the pattern.
- Wide-range used[8] for character, word, signature and handprint recognition systems.

We used Algorithm based on image processing operation Hit-Miss which removes only pixels that satisfy the template and preserve others, see figure 6.

Templates which use in the thinning algorithm:

- Templates for Hit-Miss:



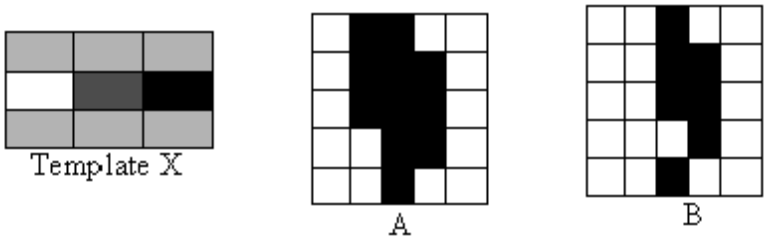
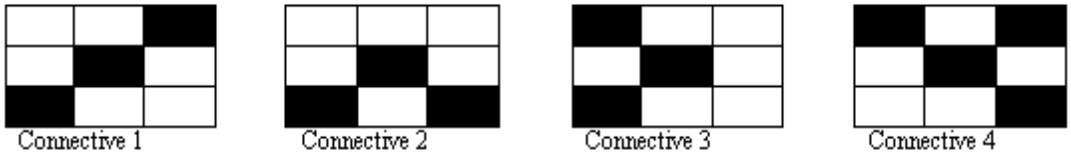
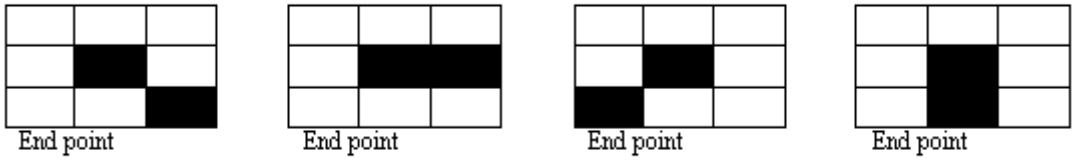


Figure 6. Image B produced after applying Hit-Miss operation over Image A using Template X once.

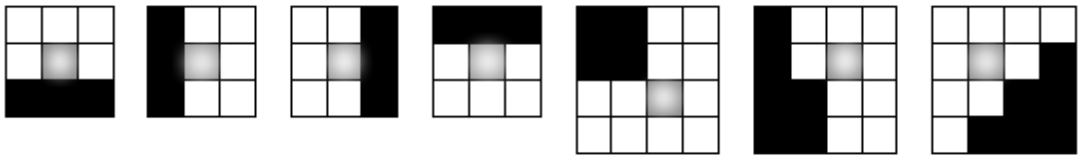
– Templates for connectivity:any pixel that connect with more than two sets of pixels.



– End point templates: any pixel that connect only with one pixel.



– Templates for noise removing:pixel needs to be removed to get smothed skeleton.



Thining Algorithm

Repeat step 1...3 until there is only end point pixels and connective pixels in the image.

– Step 1: search noises and remove them applying the templates for noise removing.
– Step 2: For each pixel from left to right:
if the template is not in the set of Connective templates and is not in the set of end point templates.
then apply hit-miss operation using template 2 or template 4.
– Step 3: For each pixel from up to down :
If the template is not in the set of Connective templates and is not in the end point templates.
then apply hit-miss operation for template 1 or template 3.

Figure 7 shows the result of the thinning algorithm and how it saved geometry and the connections between characters.

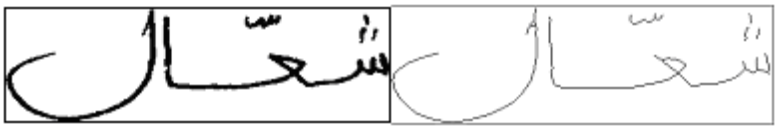


Figure 7. Arabic word before and after applied thinning algorithm

4.3. Baseline detection

Most letters in Arabic word are connected with other letter depending on letter's position (Beginning, Middle, and End) within the word ,figure 8 shows all arabic charcters forms within the word. These connections are based on line along the word which is called baseline. There is an exception where some letters in Arabic language are not connected with the next letter when lies in any position inside the word see figure 9.

Position in The word / Letters Classes		Standalone	Beginning	Middle	End
1	Class 'Alif	ا	ا		ـا
4	Class Daal	د	د		ـد
5	Class Raa'	ر	ر		ـر
6	Class Waw	و	و		ـو
2	Class Baa'	ب	بـ	ـبـ	ـب
3	Class Jeem	ج	جـ	ـجـ	ـج
7	Class Seen	س	سـ	ـسـ	ـس
8	Class Saad	ص	صـ	ـصـ	ـص
9	Class Taa'	ط	طـ	ـطـ	ـط
10	Class 'Ayn	ع	عـ	ـعـ	ـع
11	Class Faa'	ف	فـ	ـفـ	ـف
12	Class Laam	ل	لـ	ـلـ	ـل
13	Class Meem	م	مـ	ـمـ	ـم
14	Class Haa'	ه, هـ	هـ	ـهـ	ـه
15	Class Yaa'	ى	يـ	ـيـ	ـي

Figure 8. All forms of arabic charcters

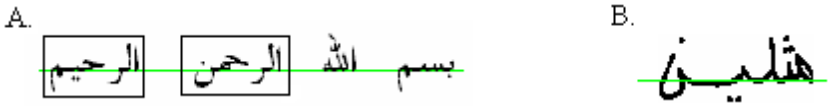


Figure 9. Sentence and handwritten word from Arabic language with baseline. Inside the rectangle one word consists of two subwords.

Some characters have ascenders which is a part of character's shape that lies above baseline, and descenders which is a part(s) letter's shape that lies under baseline, there are also special marks such dots, Hamza(zigzag shape) and Madda. All these marks lie above or under some of letters which means above and under the baseline itself. See figure 10. Due to these facts it's important to find baseline exactly and to extract correct features along the baseline's position, and over, under baseline.

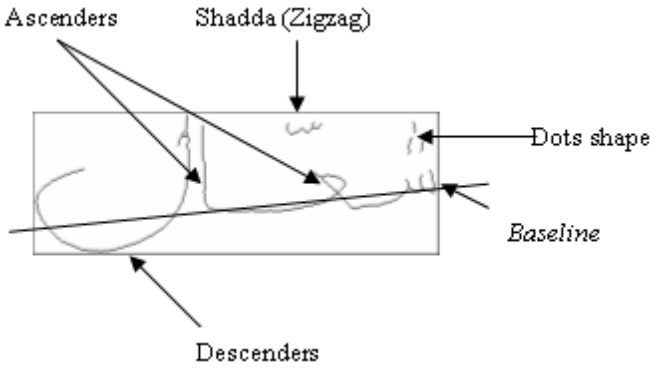


Figure 10. Shows Arabic handwritten word, its baseline, shadda, dots.

Baseline algorithm:

Assumption: we assume that baseline is rotated horizontally with some angle threshold between +20 - -20, and use the fact that maximum number of pixels are located along the baseline,figure 11 shows horizontal histogram for Arabic word.

Figure 11. Horizontal histogram.

Input: skeleton for Image's word. Output: angle, the position of exact baseline

- Step1: rotate the image to the left with some threshold and from the rotated image compute the peak of horizontal histogram.
- Step2: rotate the image to the right with some threshold and from the rotated image compute the peak of horizontal histogram.
- Step3: we compare the step1 result with step2 result which give us the indication in which direction the baseline will be.
- Step4: Repeat Rotate image with some threshold's value to the direction that identified in step3.

  Compute maximum value from the horizontal histogram.
  If maximum value $_{new}$ ¿ maximum value $_{previous}$ then
  Update the maximum value.
  Save the image $_{new}$.
  Save the angle $_{new}$.
  Else
  Increment FailedCcounter by one
- Until the FailedCounter not exceeded some threshold
- Return the angle and the index of maximum horizontal histogram for that angle.

In figure 12 there is an image of town name contains two arabic words. After applied the thinning algorithm the words converted to skeleton. Baseline detection algorithm use the skeleton as input.

The output of the above descriped algorithm shows the basline location and it's angle. Figure 13 also

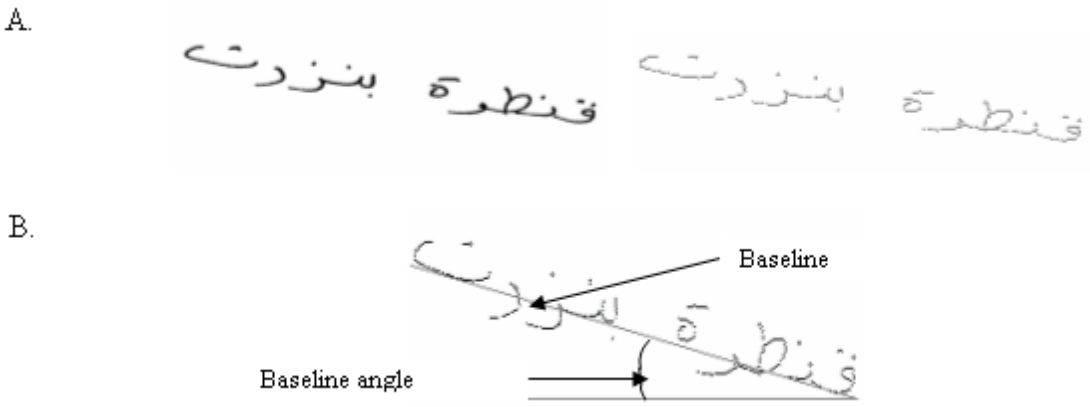


Figure 12. Example shows the input and output of the Baseline algorithm

shows how the algorithm found the correct baseline and stopped when the peak value decreased.

Rotate value (degree)	Image after rotation	Horizontal Histogram	The highest peak
-5			20
+5			32
+8			46
+12			54
+14			30

Figure 13. Shows the Baseline algorithm behavior

5. CONCLUSION AND FUTURE RESEARCH

In this paper we present a review about pattern recognition and its importance, and it's applications, also we list the characteristics of writing for Arabic language, and focused in one of important phases in recognition systems which is Preprocessing. Moreover we discussed the benefits of basline detection and it's role in feature extraction phase, also we described and implemented a skeleton algorithm and algorithm to detects and find baseline for Arabian words. We tested these algorithms using IFN/ENIT of handwritten Tunisian town names and the algorithms work properly.

In the future we will use the result from this step to extract features and organized the representation depending on skeleton and baseline information.

6. ACKNOWLEDGMENTS

i would like to present thanks to my supervisor Prof. Victor V. Alxandrov for his valuable disscussion and comments.

REFERENCES

1. Devijver, P. A. and J. Kittler (1982). Pattern Recognition: a Statistical Approach. London: Prentice Hall International.
2. Ahmad T. Al-Taani, "An Efficient Feature Extraction Algorithm for the Recognition of Handwritten Arabic Digits," INTERNATIONAL JOURNAL OF COMPUTATIONAL INTELLIGENCE, VOL. 2, 2005, ISSN: 1304–4508.
3. Koerich, A. L., R. Sabourin, C. Y. Suen, "Large Vocabulary Off-Line Handwriting Recognition: A Survey," Pattern Analysis and Applications, v. 6, no. 2, pp. 97–121, July 2003.
4. Tal Steinherz, Ehud Rivlin, Nathan Intrator, "Offline cursive script word recognition - a survey," International Journal on Documents Analysis and Recognition (IJDAR), pp. 90–110, September 1999.
5. Plamondon, Rejea and Sargur N. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 22, no. 1, January 2000.
6. M. Pechwitz, S. S. Maddouri, V. Maergner, N. Ellouze, and H. Amiri. IFN/ENIT – database of handwritten Arabic words. In Proc. of CIFED 2002, pages 129-136, Hammamet, Tunisia, October 21–23 2002.
7. PETROS A. MARAGOS, RONALD W. SCHAFER, "Morphological Skeleton Representation and Coding of Binary Images," IEEE TRANSACTIONS ON ACOUSTICS. SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-34, NO. 5, pp. 1228–1244, October 1986.
8. Lam, L., Lee, S., and Suan, C.Y., "Thinning Methodologies – A Comprehensive Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(9): pp. 869–885, September 1992.
9. M. Blumenstein, C. K. Cheng and X. Y. Liu , "NEW PREPROCESSING TECHNIQUES FOR HANDWRITTEN WORD RECOGNITION," 2nd IASTED International Conference Visualization, Imaging, and Image Processing , Benalmadena, Malaga, Spain, September 9–12, 2002.

This paper was recommended for publication by V. Venets, a member of the Editorial Board