A New System for Offline Printed Arabic Recognition for Large Vocabulary: SPARLV

Mariem Miledi Dhouib, Slim Kanoun
A *REsearch Group on Intelligent Machines (REGIM), University of Sfax, National School of Engineers (ENIS),Route Soukra Km 3.5 B.P. 1173 3038 Sfax, Tunisia*
E-mail: mariem_miledi@yahoo.fr, slim.kanoun@enis.rnu.tn

Abstract

This paper presents a contribution for the Arabic printed recognition. In fact, we are interested in off-line printed Arabic word recognition for a large vocabulary. The proposed system SPARLV uses the analytical approach through the segmentation into characters to succeed to a generation of letter hypotheses as well as word hypotheses using a lexical verification in a pre-established dictionary of the language. Our proposed system SPARLV is able to put valid hypotheses of words thanks to the lexical verification.

1. Introduction

For several decades, many researches have been undertaken in field of recognition of the Arabic script which always remains object of several studies and researches. Some researchers were interested in on-line recognition, which is generally much more effective than off-line recognition and which has samples that are much more informative than those of the off-line recognition. Actually, the field of the off-line recognition of the Arabic printed script stills an important challenge to rise and an open field because of its multiple difficulties with which the researchers are confronted in particular its semi-cursive nature and the great morphological variability of its characters. The performance of such a system of recognition is narrowly dependent on the quality of the document, on the vocabulary used as well as on the size, the font and the style of the characters. By observing more closely works of literature, we can notice that the first systems of recognition of the Arabic script were dedicated for characters recognition such as systems of [1], [2], [3], [4] and [5]. Then systems of recognition of words like systems of [6], [7], [8] and [9] appeared and little by little there exist even systems of text recognition namely systems of [10], [11] and [12]. Never the less these systems present a minority and did not solve all the problems of recognition of the Arabic script. In fact, most existing systems in the literature have treated the limited vocabulary [13], [14] and [6]. However there exist some systems which treat a wide vocabulary [10], [15] and [16] and little of them which integrate a module of lexical validation in a language dictionary [17], [18], [19] and [20]. That is why we propose in this paper a new system of recognition of printed Arabic multi-size, multi-font and multi-style words for a large vocabulary. Our proposed system SPARLV is able to put valid hypotheses of words thanks to the lexical verification in a language dictionary in a phase of post-processing.

In the next section of this paper, we will detail the system suggested for the recognition of printed Arabic words by the analytical approach while validating hypotheses of words by a lexical checking in a dictionary of language. Then we will present some experimental results. Finally, we achieve this paper by a conclusion and some prospects.

2. The proposed system for the recognition of printed Arabic words

In this section, we describe our system SPARLV by detailing its principal modules.

2.1. Pre-processing

The pre-processing has as a role to prepare the image to the treatment. Among pre-processing techniques, we quote the most important stage to know namely the binarisation which consists in reducing the quantity of information to be treated when the image is in levels of gray. Indeed, it makes it possible to recover the printed paper form by isolating it from the background. The technique used in the case of our system SPARLV is which seeks a total threshold to separate the levels of gray which correspond to the features of writing of those corresponding to the background; which allows to convert an image of levels of gray into a binary image made up of 2 values 0 and 1 (respectively representing a white pixel and a black pixel).

2.2. Detection of the diacritics

To carry out the extraction of the diacritics as well as their characteristics (coordinates, width, height, number of black pixels, number of occlusions), we start by extracting the connected masses (a connected mass can be a diacritic, an isolated character, or a piece of word) composing each word. Then, we pass to locate the diacritics and to eliminate them from the matrix of pixels after having safeguarded their coordinates and their dimensions (width, heighté). In fact, our system SPARLV detects diacritics by searching connected masses composed of few black pixels (using a predefined threshold fixed after several tests) and which do not have intersection with the ligature (connection between two characters).

2.3. Estimation of the thickness of the ligature

The thickness of the ligature (connection between two characters) is another important characteristic of a font and which distinguishes one font from another, a style from another and a size from another. And since in this work we adopt the analytical approach, we need this characteristic in the phase of segmentation. In fact, the success of the latter depends strongly on the good estimation of the thickness of the ligature value. In this current work, we estimate the thickness value as the difference between the higher frontier and the lower frontier of the ligature. More precisely and as shown in the Figure 1 the two indices row1 and row2 (of two larger absolute values of the difference between each successive values of the profile of the horizontal projection) correspond to the higher and lower frontiers of the thickness of the ligature [15]. Consequently, we have: Higher frontier = min (row1, row2) and lower frontier = max (row1, row2). These frontiers are described on the level of Figure 1.

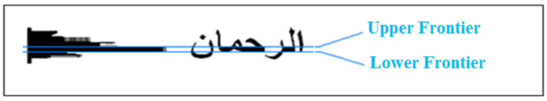


Figure 1. *Higher and lower frontiers of the thickness of the ligature*

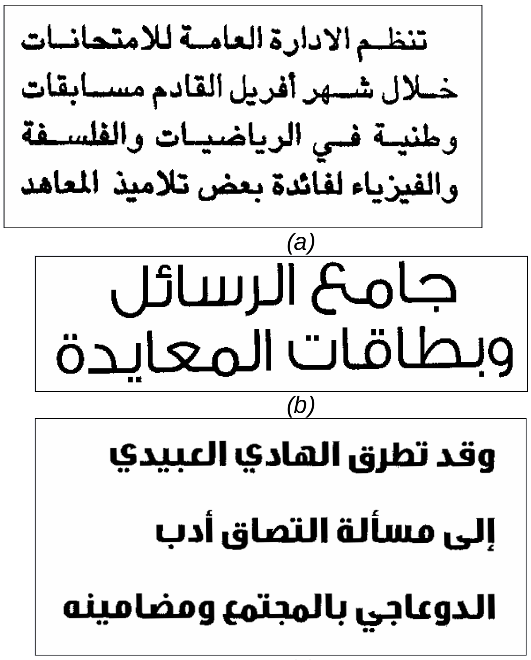
2.4. Segmentation into characters

The phase of segmentation is a too important stage in any system of recognition of the writing based on an analytical approach. Indeed, the major difficulty encountered by such an approach is to lead in a robust way to good hypotheses of segmentation. We have developed an algorithm of segmentation which is based on the analysis of the profiles of vertical and horizontal projections of the image of word as well as on the particular characteristics of the Arabic writing. More precisely, as long as the number of black pixels of the considered column is equal to the thickness of the ligature value we pass to the following column and we segment only when the number of black pixels of a column will be greater than the thickness.

2.5. Characters recognition

Following the stage of segmentation of the image of word in letters, we carry out the recognition of these letters. On the level of this section, we describe the principal stages carried out in order to make the character recognition.

2.5.1. Construction of a training database of letters. We built a training database of the Arabic shapes of letters (the shape of letter is a letter without diacritics) independent of the algorithm of segmentation used. Indeed, this database is manually elaborated by segmenting images of texts using a software of image processing. Actually, we have scanned, on a level of resolution of about 300 dpi, 200 blocks of texts taken from most known magazines and newspapers written in Arabic language. These blocks of texts cover three various Arabic fonts which are most used in the Arabic magazines and newspapers. We have chosen to build real images (extracted from a scanner) and not synthetic ones (like those used in other researches), because images to be recognized via an OCRS (Optical Character Recognition System) are images scanned from printed papers (newspapers, magazines, administrative documents,é). Figure 2 presents some images of blocks of texts scanned for each of the three fonts treated in this current work.



تنظم الادارة العامة للامتحانات
خلال شهر أفريل القادم مسابقات
وطنية في الرياضيات والفلسفة
والفيزياء لفائدة بعض تلاميذ المعاهد

(a)

جامع الرسائل
وبطاقات المعايدة

(b)

وقد تطرق الهادي العبيدي
إلى مسألة التصاق أدب
الدوعاجي بالمجتمع ومضامينه

(c)

Figure 2. *Examples of images of texts: (a) Font n°1, (b) Font n°2 and (c) Font n°3*

We have obtained a training database consisted of 6606 shapes of letters strongly similar to the segments resulting from the automatic algorithm of segmentation (used for the phase of segmentation: we adopt the same principle of the automatic segmentation in the manual segmentation). The shapes of letters constituting the database thus built are distributed between the 31 shape classes of letters which will be used by the classifier (see section 2.5.3).

2.5.2. Features Extraction. The definition of features is one of the most delicate stages in the construction of a system of recognition. Therefore, the choice of the characteristics influences the performances of such a system.

We have tested several features used in the literature such as: characteristics suggested in [21], representations in chains of code of Freeman and other characteristics. We noted that the first features gave the best results. That is why, we have used a vector of characteristics containing some characteristics, among those proposed in [21], which are the following ones: intersections with lines, profiles high, low, right and left, horizontal and vertical projections and extrema high, low, left and right. We have added to these features Zernike moments [22] in order to ameliorate the results of classification. Following this description of the characteristics, each elementary segment resulting from the phase of segmentation, in order to be classified, will be represented by a vector of features whose size is equal to 82.

2.5.3. Classification and recognition. The recognition of the cursive writing with large vocabulary undoubtedly presents one of the cases of the most extreme difficulties. In order to increase the chances of a good recognition, it is necessary to choose the most dedicated and most effective approach. It is in this direction that we have chosen the analytical approach as a method of recognition of words. Let us recall that this approach requires one local interpretation based on the segmentation of the word into elementary segments (letters or parts of letters). For the recognition of each segment, resulting from the phase of segmentation of words, we have adopted the statistical approach. In fact, we have used K Nearest Neighbors classifier (KNN) with K=1 like a method of classification in order to evaluate the probability between the vector of characteristics of the segment to be recognized and those of the models of segments stocked in the training database of shapes of letters that we have already constructed (see section 2.5.1), by calculating the Combera distance (if we have two vectors with n variables X(x1,x2,..,xn) and Y(y1,y2,..,yn) then the Combera distance is equal to ∑|xi-yi|/|xi+yi|). This classification succeeds, for each segment, with a list of proposals of classes of segments classified by an ascending order according to the calculated distances. At this level, we mention that our system emits one class of shape for each character but after adding diacritics we obtain more than one hypothesis for each character. Actually, the choice of the value of K was based on the fact of wanting to satisfy a compromise which is the following: on the one hand, if we decide to emit as hypotheses of shapes of letters more than only one, then we will be likely to have an explosion of the space of words hypotheses. On the other hand, the fact of taking a value of K higher than 1 will guarantee the emission of the good shape class of letters even if this class did not appear in first position among the closest neighbors classes. In our current research, we have adopted the KNN classifier which uses the vector of features described previously containing 82 characteristics (see section 2.5.1). Once we have found the class of the segment to be recognized, it is enough to add the diacritics to the shape of the letter corresponding to the found class in order to generate hypotheses of letters corresponding to the considered segment.

2.6. Constitution of words hypotheses by concatenation of characters hypotheses

As we have already announced, our system allows the recognition of a large vocabulary. However, according to [23] and [24] almost total of the vocabulary of the Arabic language with everyday use consists of a root to which a formed affixal combination of a prefix, an infix or a suffix is added. This major part of the Arabic vocabulary subjugated by decomposition in affixes and roots constitutes the decomposable vocabulary. It is in this direction that we were interested to recognize decomposable words. To ensure this recognition, we have adopted the analytical approach using a language dictionary in order to ensure the lexical checking of the emitted words hypotheses. Our system SPARLV starts by generating words hypotheses. In fact, after having emitted hypotheses of letters following the stage of characters recognition, our system generates hypotheses of words by concatenation of the various hypotheses of letters. The emission of these various hypotheses of words enabled us to calculate a first rate of recognition of words which was improved thanks to the lexical checking in a language dictionary (the rate of recognition thus obtained is clarified on the level of section 3 of this paper).

2.7. Lexical checking by using a language dictionary

Considering that the phase of segmentation produces over-segmented letters and even the classifier can sometimes generate erroneous hypotheses of letters, we propose, with an aim of improving the recognition of the words containing an over-segmentation or an error of characters recognition, to use a lexical validation of the hypotheses words in a dictionary of decomposable words.

2.7.1. Elaboration of the dictionary. In this work, we have elaborated a dictionary of decomposable words to be able to show the importance of the lexical validation in a language dictionary. Indeed, starting from a lexicon of 472 schemes and 100 healthy and tri-consonant Arabic roots our dictionary contained, in the first time, 47200 Arabic decomposable words. In fact, starting from the lexicon of schemes and an algorithm which substitutes the letters ñFaò, ñAinò and ñLamò by the three letters of each of the 100 roots, this first version of our dictionary was elaborated.

It is important to mention that the Arabic roots do not join systematically with all the coherent affixal combinations of the language. Indeed, with each root corresponds a well defined list of affixal combinations and consequently a list of schemes with which it can be joined. That is why, we have obtained a dictionary containing only 19034 words after having carried out a semantic filtering of the first version of our dictionary (manually using Windows XP) while leaving only decomposable words resulting from the schemes which are appropriate for each of the 100 used roots. At this level, the analysis of this dictionary shows the variation amongst words accepted by the language of a root to another. This number varies from 335 for the root ñ قطعò up to 48 for the root ñ ثلج ò. We did not extract other notes considering that the number of used roots is relatively limited by report with the enormous number of tri-consonant roots of the Arabic language

.

2.7.2. Lexical checking of words hypotheses in the dictionary. After having generated hypotheses of words by the concatenation of the recognized letters, a lexical checking (in the dictionary described above) of each hypothesis of word is carried out in order to keep only hypotheses of words accepted by the language and to reduce consequently the number of hypotheses resulting from our system with an aim of converging towards the good hypothesis and of improving the rate of recognition obtained by our system. On the one hand, this lexical checking consists in checking if the hypotheses of words, emitted by our engine of recognition, belong to the dictionary. More precisely, if all letters of the word to be recognized are correctly recognized by our engine of characters recognition, then the good hypothesis of word will certainly appear among those generated by our engine of recognition of words and consequently found in our dictionary. Then other hypotheses which are not found in the dictionary will be removed from the final list of the hypotheses of the word to be recognized. In addition, if there were an error of recognition caused by an error of segmentation or by a bad classification of a letter constituting the word to be recognized, it is possible not to find any hypothesis confirmed in the dictionary. Then, when it was about only one error of recognition, we proceed by removing a letter of the word hypothesis (among those emitted by our engine of recognition of words) ñhypo_iniò to lead to a new hypothesis ñhypo_secò which corresponds to the remainder of the hypothesis initially put by our engine of recognition. Then, we carry out a second research, which allows extracting from the dictionary other hypotheses including in their character string (while respecting the order of the letters) the whole of the characters of the new hypothesis ñhypo_secò. By this way, we can then generate other hypotheses from hypotheses already rejected by the lexical checking in our dictionary. Thus, by applying this method, we could succeed the recognition of words containing an error of recognition and consequently obtain more interesting rate of words recognition.

3. Experimental results

In order to evaluate the performance of our system, we have tested our system SPARLV on a test database made up of 100 images of Arabic decomposable words scanned on a level of resolution of about 300 dpi, from magazines and newspapers written in Arabic. The images of words constituting the test database cover various fonts, sizes and styles. Moreover, words composing this test database have variable lengths and resulting from different Arabic roots. We have obtained a rate of words recognition equal to 84%. This rate is interesting considering the diversity of the treated fonts and styles and the fact that we used scanned images (and not synthetic ones). It is also important to mention that this rate was improved thanks to the integration of the module of post-processing which allows the lexical validation of the words hypotheses emitted by our system as well as the generation of valid (lexically) words hypotheses even if the hypothesis has an error of recognition of a letter. This while guaranteeing the appearance of the good hypothesis among the generated hypotheses thanks to the lexical checking in the language dictionary.

4. Conclusion and prospects

In this paper, we have proposed a new system of recognition of printed Arabic words for large vocabulary SPARLV by analytical approach and using a language dictionary in phase of post-processing. The proposed system SPARLV allows the lexical checking of the emitted words hypotheses and thus the important reduction of their number. Moreover, we have proved the importance of the lexical checking in the improvement of the rate of recognition obtained by our system even if it has a letter which was not correctly recognized. The effectiveness of the integration of such a module of post-processing is very encouraging what leaves the field open to several perspectives. In fact, our system SPARLV can lead to rates of recognition much more interesting by improving its robustness by the use of more powerful tools for classification (than the KNN classifier) such as the neural networks, the hidden models of Markov and the amelioration of the dictionary as well as the databases used by a more large number of examples. In [25] authors showed that, in the case of the mono-font, the affixal approach proposed in [20] is much more powerful (in term of the number of emitted words hypotheses and execution time) than analytical approach using a language dictionary in post-processing. As a major perspective, we aim to comparing the results of our system SPARLV with those given by the affixal approach in the case of multi-font and multi-style Arabic scripts.

References

[1] Bushofa B. M. F., M. Spann, « Segmentation and recognition of Arabic characters by structural classification », *Image and Vision Computing IVC 97*, vol. 15, n° 3, 1997, p. 167-179.

[2] M. Fakir, M. M. Hassani and C. Sodeyama, « Recognition of arabic characters : an alternative approach », *Proc. Of International Conference on Image and Signal Processing ICISP 01*, Agadir, 3-5 may 2001, Maroc, p. 522-529.

[3] Khorsheed M.S., « Off-Line Arabic Character Recognition ï A Review », *Pattern Analysis and applications PAA 02*, vol. 5, n° 1, 2002, p. 31-45.

[4] Zheng L., « Machine printed Arabic Character Recognition Using S-GCM », *International Conference on Pattern Recognition ICPR 06*, 20-24 august 2006, p. 893-896.

[5] Ben Amor N., Ben Amara N., « An approach for Multifont Arabic Characters features Extraction based on Contourlet Transform », *International Conference on Documents Analysis and Recognition ICDAR 07*, Parana, 23-26 september 2007, Brazil, p. 1048-1052.

[6] Khorsheed M.S., Cloksin W.F, « Spectral features for Arabic word recognition », *International Conference on Acoustics, Speech, and Signal Processing ICASSP 00*, vol. 6, 2000, p. 3574-3577.

[7] Sari T., Sellami M., « Deux méthodes morpho-lexicales pour la correction des mots Arabes issus des systèmes OCR », *African conference on Research in Data processing and mathematics applied CARI 02*, 2002.

[8] Broumandnia A., Shanbehzadeh J., Nourani M., « Segmentation of Printed Farsi/Arabic Words », *International Conference on Computational Science and its Applications ICCSA 07*, 13-16 may 2007, p. 761-766.

[9] Ben Cheikh I., Belaïd A., Kacem A., « A Novel Approach for the Recognition of a Wide Arabic Handwritten Word Lexicon », *International Conference on Pattern Recognition ICPR 08*, Florida, 8-11 december 2008, USA, p. 1-4.

[10] Azmi R., Kabir E., « A new segmentation technique for omnifont Farsi text », *Pattern Recognition Letters PRL 01*, vol. 22, n° 2, 2001, p. 97-104.

[11] Khorsheed M.S., « Off-Line recognition of omnifont Arabic text using the HMM Toolkit (HTK) », *Pattern Recognition Letters PRL 07*, vol. 28, n° 12, 2007, p. 1563-1571.

[12] Al-Muhtaseb A., Mahmoud S., Qahwaji R., « Recognition of off-line printed Arabic text using Hidden Markov Models », *Signal Processing SP 08*, vol. 88, n° 12, *2008*, p. 2902-2912.

[13] Amin A., Mansoor W., « Recognition of printed Arabic text using neural networks », *International Conference on Documents Analysis and Recognition ICDAR 97*, 18- 20 august 1997, Germany, p. 612-615.

[14] Ben Amara N., Belaïd A., Ellouze N., « Modélisation Pseudo Bidimensionnelle pour la reconnaissance de chaînes de caractères Arabes imprimés », *CIFED 98*, Quebec, 11-13 may 1998, Canada, p. 131-140.

[15] Zheng L., Hassin Abbas H., Tang X., « A new algorithm for machine printed Arabic character segmentation », *Pattern Recognition Letters PRL 04,* vol. 25, 2004, p. 1723-1729.

[16] Touj S., Ben Amara N., Amiri H., « Two approaches for Arabic Script recognition-based segmentation using the Hough Transform», *International Conference on Documents Analysis and Recognition ICDAR 07*, Parana, 23-26 september 2007, Brazil**,** p. 654-658.

[17] Amin A., Mari J.F, « Machine Recognition and Correction of Printed Arabic Text », *IEEE-Transactions on Systems, Man, and Cybernetics SMC 89*, vol. 19, n° 5, 1989, p. 1300-1306.

[18] Amin A., Al-Fedaghi S., « Machine recognition of printed Arabic text utilizing natural language morphology », *IEEE-Transactions on Systems, Man, and Cybernetics SMC 91*, vol. 35, n° 6, 1991, p. 769-788.

[19] Goraine H., Usher M., Al-Emami S., « Off-Line Arabic Character Recognition », *IEEE Computer Magazine ICM 92*, 1992.

[20] Kanoun S., Identification et Analyse de Textes Arabes par Approche Affixale, Doctorat Thesis, University of Rouen, 2002.

[21] Heutte L., Reconnaissance de caractères manuscrits : Application à la lecture automatique des chèques et des enveloppes postales, Doctorat Thesis, University of Rouen, december 1994.

[22] Zernike F., « Diffraction theory of the cut procedure and its improved form, the phase contrast method », *Physica 34*, Vol. 1, 1934, pp. 689-704.

[23] Ben Hamadou A., Vérification et Correction Automatiques par Analyse Affixale des Textes Ecrits en Langage Naturel : le cas de lôArabe non Voyellé, Doctorat Thesis, University of Sciences, Technology and Medicine of Tunis, 1993.

[24] Ammar S., Dichy J., *Al-Chamil fi tasrif al-afal*, Paris, Bescherelle Collection, 1999.

[25] Kanoun S., Slimane F., Guesmi H., Ingold R, Alimi A.M., Hennebert J., « Affixal Approach versus Analytical Approach for Off-Line Arabic Decomposable Vocabulary Recognition », *International Conference on Documents Analysis and Recognition ICDAR 09,* Barcelona, 26-29 july 2009, Spain**,** p. 661-665.