A New Method for Ranking of Word Hypotheses generated from OCR: The Application on the Arabic Word Recognition

Houda Gaddour [1] Hanène Guesmi [1] Fouad Slimane[1,2] Slim Kanoun [1] Jean Hennebert[2,3]
[1]ENIS, BP 1173, 3038, Sfax, Tunisia
[2]DIVA Group, University of Fribourg, Bd de Pérolles 90, CH-1700 Fribourg, Switzerland
[3]Computer Science Department, EIAFR, HES-SO // Fribourg, Switzerland
E-mail: houda.gaddour@yahoo.fr, fouad.slimane@unifr.ch, slim.kanoun@gmail.fr, Jean.Hennebert@hefr.ch

Abstract

In this paper, we propose a new method for the best ranking of OCR word hypotheses in order to increase the chances that the correct hypothesis will be ranked in the first position. This method is based on the images construction of the OCR word hypotheses and the calculation of the dissimilarity scores between these last constructed images and the image to recognize. To evaluate the new proposed method, we compare them with a classic method which is based on the ranking of OCR word hypotheses under the recognition process. The experimental results of these two methods on the database of 1000 word images show that the new proposed method led to the best ranking of OCR word hypotheses.

1. Introduction

The complexity of the different scripts, the deterioration of the quality of the document images, the low resolution of the document images and the noises of acquisition induced to several ambiguities in word and text recognition process. Thus, several word hypotheses belonging to the language can be proposed. Among these hypotheses the correct hypothesis for a given word is not always placed at the first position. Consequently, two methods are proposed for the best ranking of OCR word hypotheses. The first method is based on the score estimation for each OCR word hypothesis during the post-recognition phase. The estimated score for each OCR word hypothesis is the minimal edition of the distance calculated between a given OCR word hypothesis and the words of language dictionary [1] [2]. The second method is based on score estimation for each OCR word hypothesis under the recognition process. In the framework of this second method, the OCR word hypothesis score is estimated by using a statistical model of the language [3] [4] [5] [10] or a minimal edition distance between the segmented characters resulting from a word to be recognized and the character models of training database [6].

It appears clearly that the first method presents a major problem which is the edition of the distance calculated on all the words especially in the case of an open vocabulary. Such an edition is extremely expensive in time calculation terms. The second method involves a statistical skew during the word recognition process. However, this alternative requires a continuous adjustment of the transition probabilities of letter successions which is not practical.

In this paper, we propose a new method for the best ranking of OCR word hypotheses. The proposed method is not depending on a language dictionary or on a statistical model of the language. The fundamental idea of the new proposed method is to build the images of the OCR word hypotheses and to calculate the dissimilarity scores between these last constructed images and the image to recognize. To evaluate the new proposed method, we compare them with a classic method which is based on the ranking of OCR word hypotheses under the recognition process.

In the following, we start to present the used OCR. We describe then the proposed ranking method based on the construction of OCR word hypotheses images. We detail after the developed ranking method under the recognition process. Next, we present the experimental results of these two methods on the database of 1000 word images. Finally, we achieve this paper by a conclusion and the future works.

2. Built OCR

In the framework of the development of the OCR hypotheses ranking methods evoked above, we are build a simple OCR for Arabic word recognition using the analytical approach. The steps of built OCR are [7]:

1. Word image segmentation in elementary

segments using the binding detection between letters through a vertical projection profile analysis.

2. 7 Invariant moments of Hu [7] extraction for each elementary segment from the word image segmentation
3. Elementary segments recognition. In this framework, we use a training database containing 2400 prototypes of elementary segments (150 prototypes for each shape) from the segmentation of a data set of 1000 word images scanned on a 300 dpi resolution and covering the Arabic alphabet letter forms (isolated, initial, medial and final). This data set is also in printed nature with Arabic Transparent font and with several sizes. We note in this context that after the analysis of segmentation result of the words data set cited above shows the identification of 28 elementary segment shapes. Each elementary segment from our training data set is represented by the seven invariant moments [7]. To recognize a given elementary segment, we use the simple k nearest neighbor classifier [7] and a Canberra distance as dissimilarity distance to identify the suitable shape.
4. Letter shapes recognition by elementary segments concatenation. A letter shape can be constructed by one or two or three elementary segment.
5. Letter hypotheses generation by the association of diacritic points to the letter shapes.
6. Word hypotheses generation by the exploration of the all possibility of letter hypotheses concatenation.
7. Word hypotheses filtering by a lexical verification using a language dictionary. This filtering keeps only the word hypotheses belonging to the language.

Let us note that our built OCR does not comprise any ranking method. Thus, a correct word hypothesis does not appear always at the first position. That depends on the appearance position of the correct elementary segment resulting from KNN classifier.

3. The Proposed Ranking Method Based on the Construction of OCR Word Hypotheses Images

In this section, we present a detailed description of the proposed ranking method as well as an illustration of this method on a word example.

3.1 Detailed description

The basic idea of the proposed method is to constitute the images of word hypotheses by the inverse direction of the recognition process (figure 1). Such an image of the word hypothesis is compared with the image of the initial word (word to be recognized) in order to extract a dissimilarity distance. This process will be repeated for all word hypotheses generated by our used OCR in order to have the hypotheses ranking. This ranking is realized by the ascending order of dissimilarity distances. Consequently, the correct hypothesis will be ranked at the first position with a minimum dissimilarity distance. The fundamental idea of the proposed method is that more a such word hypothesis contains less erroneous letters (because of confusion case resulting from the elementary segments recognition) more its dissimilarity distance with the word to be recognized is small.

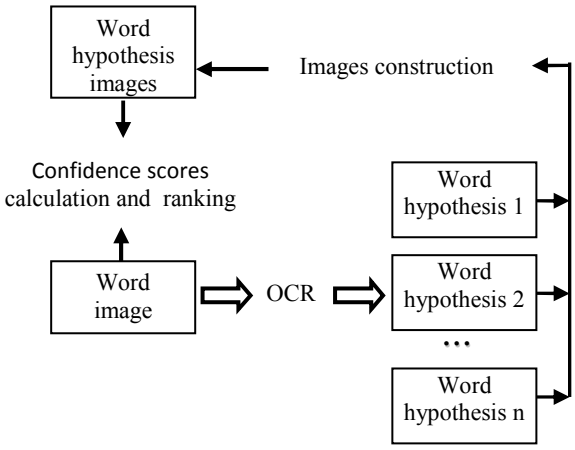


Figure 1. Synoptic diagram for the proposed method

In order to validate the contribution of the ranking proposed method, we integrate the following process in the developed OCR evoked in section 2:

1. OCR word decomposition in letters to be able to identify thereafter for each letter its corresponding shape.
2. Letter shape identification for each letter constituting the word hypothesis.
3. Elementary segment classes identification for each letter shape.
4. Random choice of 7 invariant moments of Hu [7] for each identified elementary segment from a training database used by our developed OCR. In this step, each OCR word hypothesis is represented by the succession of 7 invariant moments. Each 7 invariant moments corresponds to an elementary segment image can be constitute the OCR word hypothesis image.
5. Confidence scores calculation between the word image to be recognized and the all built OCR word hypothesis images. Such a

confidence score is calculated by the product of the calculated dissimilarity distances between the 7 invariant moments of each elementary segment constituting the word image to be recognized and the 7 invariant moments of each elementary segment constituting the OCR word hypothesis image.

6. Word hypotheses ranking by the decreasing order of calculated confidence scores.

3.2 Illustration on a word example

The recognition of the word صغير with our developed OCR generates six hypotheses without any confidence score. These hypotheses are displayed in this order: **صغير**, **صفير**, **ضمير**, **صحب**, **ضمن**. To more explain our proposed method; we present in figure 2 a synoptic diagram to illustrate the steps used by our developed OCR for the recognition of the word **(صغير)** and in figure 3 a synoptic diagram to illustrate the OCR word hypotheses ranking method for the word hypothesis **(صفير).**

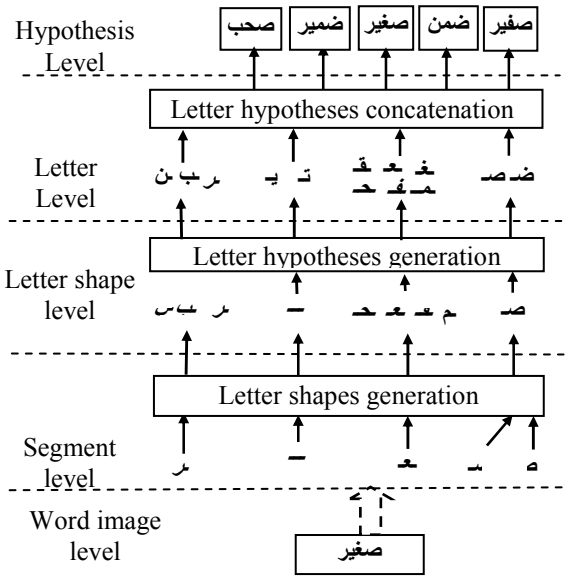


Figure 2. Synoptic diagram to illustrate the steps used by our developed OCR for the recognition of the word (صغير)

The process presented in figure 3 is repeated for the five remained hypotheses: **(ضمير)** **(صغير)** **(ضمن)** and **(صحب).** In figure 4, we present the illustration of the confidence scores calculation for all word hypotheses generated by our developed OCR after the recognition of the word **(صغير)**. We note that the erroneous elementary segments are put in rectangles.

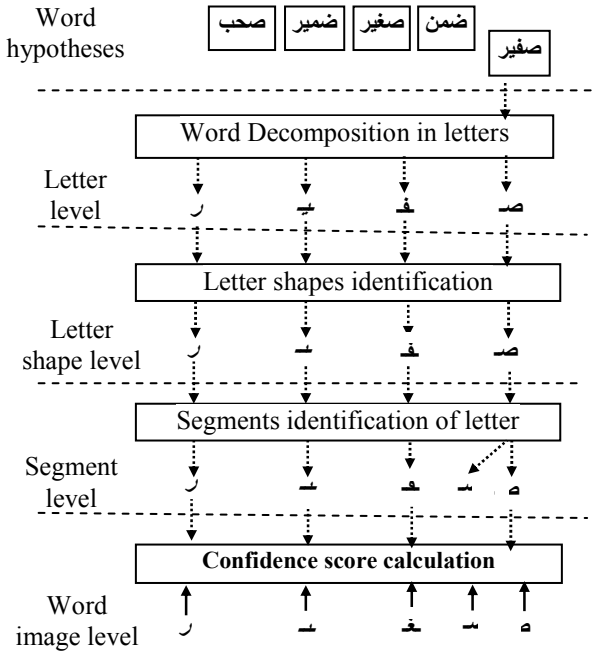


Figure 3. Synoptic diagram to illustrate the OCR word hypotheses ranking method for the word hypothesis (صفير)

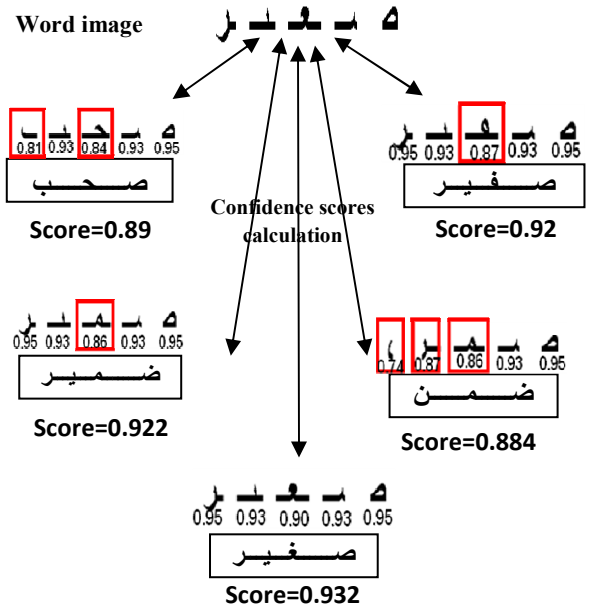
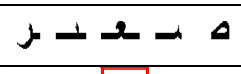
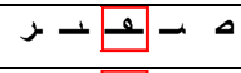
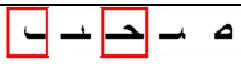
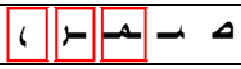


Figure 4. The illustration of the confidence scores calculation for all word hypotheses generated by our developed OCR after the recognition of the word (صغير).

At the output of the system, each word hypothesis has a confidence score. These hypotheses are ranked

by ascending order of their confidence score. After this step, the hypotheses are displayed in order presented in the following table:

Table 1. List of the words hypotheses ordered according to their scores

Confidence score	Word hypotheses images
0.932	
0.926	
0.924	
0.892	
0.870	

4. The OCR Word Hypotheses Ranking Method under the Recognition Process

To emphasize the new proposed method, we try to compare it with another. In this direction, we integrate the word hypotheses ranking method under the recognition process in the built OCR [8]. The basic idea of this ranking is to exploit the calculated dissimilarity distances between the elementary segments from word image segmentation and the elementary segments from the training database used by our built OCR. For each elementary segment constituting the word image to be recognized, the 10 nearest neighbour elementary segments are sought using KNN classifier. Thus, a confidence score is calculated for each OCR word hypothesis on the basis on the product of the calculated dissimilarity distances for the elementary segments constituting its image.

5. Experimental Results

In this section, we present the experimentation of the proposed method and the word hypotheses ranking method under the recognition process on the database of 1000 word images (others that those used for the constitution of the elementary segment images database) in printed nature with Arabic Transparent font and with several sizes. This data set is scanned on a 300-dpi resolution. In the following and to emphasize the contribution of the proposed method, we present the obtained word recognition rates for three versions of our developed OCR: a first version without any word hypotheses ranking method (called OCR 1), a second version integrating the word hypotheses ranking method under the recognition process (called OCR 2) and the third version integrating the proposed method (called OCR 3).

Table 2. Word recognition rates obtained by OCR 1, OCR 2 and OCR 3

	OCR 1	**OCR 2**	**OCR 3**
Top-1	38.98%	66.97%	89.54%
Top-2	55.15%	82.02%	95.15%
Top-3	62.92%	86.46%	97.27%
Top-4	71.71%	88.99%	98.99%
Top-5	75.55%	92.63%	100%
Top-6	79.09%	95.96%	100 %
Top-7	83.13%	97.47%	100 %
Top-8	85.15%	98.38%	100 %
Top-9	86.86%	99.19%	100 %
Top-10	89.09%	100%	100 %

The analysis of the word recognition rates presented table 2 shows that the OCR word hypotheses ranking obtained by the proposed method is more reliable than that obtained by the method under the recognition process. The word recognition rate is about 100 % at top 5 for the proposed method whereas it is about 100 % at top 100 % for the method under the recognition process.

6. The Ranking Methods Integration in HMM OCR Based Recognition System on the Large APTI Database

To generalize this method, we are concentrated now to test it in some parts of the huge APTI (Arabic Printed Text Image)[1] database with a new powerful multi-font Arabic HMM based system.

6.1 APTI Database

Available from July 2009, APTI is freely distributed to the scientific community for benchmarking purposes. At the time of writing this paper, 15 research groups have started using the APTI database. The APTI database was created in ultra low-resolution "72 dpi" with a lexicon of 113,284 different decomposable and non-decomposable Arabic words, 10 fonts, 4 styles and 10 different sizes. It contains more than 45 million Arabic word images representing more than 250 million different character shapes. Each word image in the APTI database is fully described using an XML file

[1] http://diuf.unifr.ch/diva/APTI

containing ground truth information about the sequence of characters as well as information about its generation. All Arabic letters have been adequately represented in the database. APTI is divided into 6 sets, 5 of which are freely available to the scientific community. The sets have been designed so that the number of words and representations of letters are very close from set to set (for more details about data dispersion we refer [11]).

6.2 HMM OCR based recognition system

Our word recognition system is based on HMMs. The system used in this paper has a similar architecture to the one presented in [10]. One of its main characteristics is to be open vocabulary, i.e. able to recognize any Arabic printed word. Most of the system is built using the HTK toolkit. The system is working in two phases: training and recognition. These two phases are sharing the same feature extraction frontend.

In the feature extraction part, words of each image are transformed into a sequence of feature vectors computed from a narrow analysis window sliding from right to left on the word.

During training time, the Expectation-Maximization (EM) algorithm is used to iteratively refines the component weights, means and variances to monotonically increase the likelihood of the training feature.

At recognition time, an ergodic HMM is built from all character models. The features from an input image are used to compute the best state sequence in this model using a standard Viterbi decoding procedure. For more details, we refer to [12].

To use the presented method in the post-processing step of our HMM based system, the idea consist of :

- The image of each hypothesis building using the proposed method in [11];
- Features extraction for the built hypothesis images;
- Confidence scores calculation between the feature extraction of word image to be recognized and the feature extraction of all built OCR word hypothesis images;
- Word hypotheses ranking by the decreasing order of calculated confidence scores.

7. Conclusion

In this paper, we are presented a new method for OCR word hypotheses ranking based on the images construction of hypotheses generated from OCR for a word to be recognized. This proposed method is experimented and is compared with an another method OCR word ranking under the recognition process on the basis of a database of 1000 word images scanned on a 300 dpi resolution. The obtained results confirm the originality contribution of the proposed method for a best ranking of OCR word hypotheses. The correct hypothesis is ranked at the first or at the second position for the majority of the 1000 word images of our database. Several experimentation works of the proposed method on a representative, a very big size and with low resolution word database (APTI) [11] is in hand. Also, several others integration works of the proposed method on the HMM OCR based recognition system [10][12].

References

[1] S Carbonnel, “Integration and modeling of language skills for on-line handwriting recognition*”, PHD thesis INSA de Rennes, France*, 2005.

[2] J. F. Pitrelli, J. Subrahmonia and M. P. Perrone. “Confidence modeling for handwriting recognition: algorithms and applications”, *IJDAR,* 2006.

[3] M. S. Khorsheed, “Recognising handwritten Arabic manuscripts using a single hidden Markov Model”, *PRL*, 24(14), pp. 2235-2242, 2003.

[4] S. Quiniou, “Integration of language skills for on-line recognition of handwritten texts”, *PHD thesis, University of Renne*, 2007.

[5] M. Zimmermann, J. C. Chappelier and H. Bunke “Offline Grammar-Based Recognition of Handwritten Sentences”. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 28 (5), pp. 818-821, 2006.

[6] G.Kim, Venu Govindaraju, Sargur N Srihari, "An architecture for handwritten text recognition systems", *IJDAR*, 1999.

[7] S. Kanoun, A. Alimi, Y. Lecourtier, “Natural Language Morphology Integration in Off-Line Arabic Optical Text Recognition”, *IEEE Transaction on Systems, Man and Cybernetics – Part B: Cybernetics, Vol 41, NO. 2,* pp. 579 – 590, 2011.

[8] H. Guesmi, “OCR Word hypotheses Ranking Methods Study and Design of incremental training for Arabic Printed mono-font and multi-size Text Recognition Using Affixal Decomposition”, *Master's thesis, University of Sfax*, 2007.

[9] H. Gaddour, “Optimization of the decision and the learning of a words and texts recognition system”, *Master's thesis, University of Sfax,* Tunisia, 2005.

[10] F. Slimane, R. Ingold, A. M. Alimi, and J. Hennebert, “Duration models for arabic text recognition using hidden markov models,” *CIMCA*, pp. 838–843, Vienna, Austria, 2008.

[11] F. Slimane, R. Ingold, S. Kanoun, A. M. Alimi, and J. Hennebert. “A New Arabic Printed Text Image Database and Evaluation Protocols”. *ICDAR*, pp. 946-950, Barcelona, Spain, 2009.

[12] F. Slimane, R. Ingold, S. Kanoun, A. M. Alimi, and J. Hennebert, “Impact of character models choice on arabic text recognition performance,” *ICFHR*, pp. 670–675, 2010.