Script Identification for Arabic and Latin, Printed and Handwritten Documents

Slim KANOUN*, Ikram MOALLA*, Abdel ENNAJI*, Adel M. ALIMI**
* Perception, System, Information Laboratory (PSI),
University of Rouen
76821 Mont Saint Aignan Cedex, France
E-mail : {slim.kanoun, abdel.ennaji,ikram.moalla}@univ-rouen.fr

**REsearch Group on Intelligent Machines (REGIM)
University of Sfax, ENIS, DGE
BP. W-3038 - Sfax – Tunisia
E-mail : adel.alimi@ieee.org

1. Introduction

Current research in the field of script and document analysis and their recognition aims at conceiving and establishing an automatic system able to discriminate a certain number of scripts in order to select the appropriate recognition system to a given document. Nowadays, the daily produced documents worldwide can contain printed and handwritten zones of different scripts. Consequently, the automation script identification and printed or handwritten nature of the considered blocks has became a necessity in any text recognition system.

Most research works in the field of identification of the script shows that the majority of them concern only the identification of scripts in printed documents, except Hochberg et al [Hoc99] proposed a method for the identification of handwritten documents.

The all proposed methods in the literature, dealing with the case of printed script, can be classified in three main categories according to the entity used to carry out script identification. One can distinguish methods based on the analysis of blocks of text, methods based on the analysis of lines and methods based on the analysis of connected components.

In the first case, a text block is considered as a whole. Such a method consider a text block as being only one entity and thus does not resort to other analyses related to the text lines or connected components [Woo95] [Tan97]. The method of Wood et al [Woo95] is based on the analysis of the profiles of horizontal and vertical projections of text lines. In Tan [Tan97] authors uses the analysis and the classification of textures of uniform text blocks (standardized line spaces and inter-word spaces).

The methods based on the analysis of text lines are focused on the analysis of the morphological features of the texts in different scripts. The extraction of features is made at the beginning on the connected components, and then it is generalized on the text line. The final decision is made on a certain number of text lines made up of a rather significant number of connected components [Spi97][Sue97][Lee96].

Finally, the methods based on the analysis of connected components are divided into two groups. The first one relates to the methods seeking to identify scripts by comparison either with models pre-established in the case of the printed document [Hoc97] or with prototypes of features in the case of the handwritten document [Hoc98]. The latter are often obtained starting from preliminary training set. The second category pertains to systems based on an on-segmentation of the connected components (characters, graphemes) that are used initially for the identification of script and then for their recognition [Pal97] [Abd89][Lee95].

2. Our identification approach

Most of script identification methods presented in section 1 shows that the decision is based:

- on a global analysis (on text block level) without resort to other local analysis (on text line and connected component level);
- on a local analysis (on text line and connected component levels) without having a global vision of the textual entity;
- on comparisons with pre-established models of connected components;

Moreover, the majority of these methods have dealt with the problem of identification of script only in printed documents and not in handwritten documents (the mixed documents such as forms or others are thus not considered).

In this paper, we propose an approach that relates to the identification of Arabic and Latin texts. In addition, this method is able to detect the handwritten or printed nature of the document. The ensuing difficulty comes mainly from the cursive nature of the Arabic script and handwritten Latin.

Our objective is to conceive and establish a dynamic system able to adapt it self to a rich and homogeneous textual entity as well as to a reduced entity (only some words) as in administrative applications, bilingual forms, heterogeneous texts, etc. As a consequence of this approach, the architecture of the proposed system is based on extractors of features developed in order to account for the features of each script and to provide a vector of description the text blocks allowing their identification.

However, the system that we want to set up must be able to use all information available in textual entities to accelerate processing and to lead to a reliable script identification and its nature. Thus, our identification approach comprises two different analyses. The first one is a morphological analysis and is performed on the text block level. The second one is geometrical and concern the text line level and the connected components level.

2.1 Morphological analysis

Our morphological analysis is a global characterization of the textual entities regardless of its physical structure in terms of text lines and connected components. This analysis is supposed to extract the features from each script in terms of intrinsic morphological features. These features are :

- ***number and position of diacritics:*** a printed or handwritten Arabic text is generally rich in diacritic and dots in particularly. This is justified by the fact that several Arabic letters are based on the same shape and differs only by the number and the position of dots (bottom or up), when Latin text contains a lower number of diacritic. There are only the two letters “i” and “j” that have only one point above and there are no diacritic in bottom in a Latin text;
- ***number of occlusions:*** occlusions are more numerous in an Arabic text than in a Latin due of his cursive nature furthermore this feature is true for printed texts compared to handwritten texts owing to the fact that in certain styles of handwritings, the script writers tend not to buckle occlusions;
- ***number of character alif:*** alif is a particular character in Arabic. It generally has the shape of a vertical stroke and contains neither occlusion, nor curve. Moreover, it can be dependent neither on the right nor on the left. This character for the Arabic words materializes the articles what reflects its frequency in the Arabic texts.

2.2 Geometrical analysis

The geometrical analysis is a local characterization reflecting the physical structure of the textual entity. This stage consists of various geometrical measurements (density of pixels, eccentricity, sphericity) on the text lines and connected components. These measurements are then generalized on the text block in term of average and standard deviation. With our identification approach, one consider that the more the availability of textual entity is rich in its contents the more our system has useful information for the identification.

Among the geometrical features used within the framework we used:

• eccentricity of the connected component: ratio of the width / the height of the connected component,

• density of black pixels of the connected component compared to its bounding box : ratio of the number of black pixels / the surface of the connected component,

• density of pixels of the contour of the connected component compared to its bounding box : ratio of the number of pixels of contour / the surface of the connected component,

• sphericity of the connected component: the square of the ratio of the number of black pixels / the number of pixels of the contour of the connected component [4].

• eccentricity of occlusion: ratio of the width / the height of occlusion,

• density of pixels of the contour of occlusion compared to its bounding box : ratio of the number of pixels of contour / the surface of occlusion.

The heights of the text lines in a handwritten text depend on the style of writing of each writer. Thus, these heights vary from a handwritten text to another. On the other hand, it keeps certain stability in printed texts. For this reason, we integrate the heights of the text lines as a measure in our system.

Finally, we illustrate the morphological features used within the framework of our identification approach on a printed and handwritten Arabic text, printed and handwritten Latin text from our data set (see figures 1-4). Therefore, figure 5 summarizes the diagram of our identification approach.

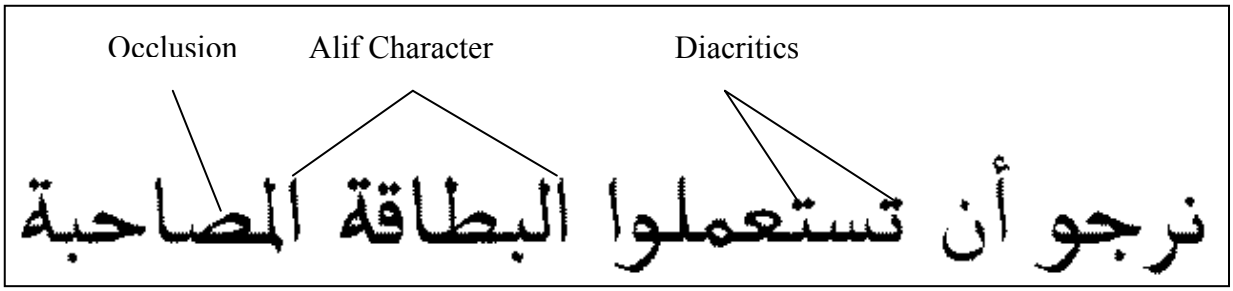

Figure 1: An example of printed Arabic text

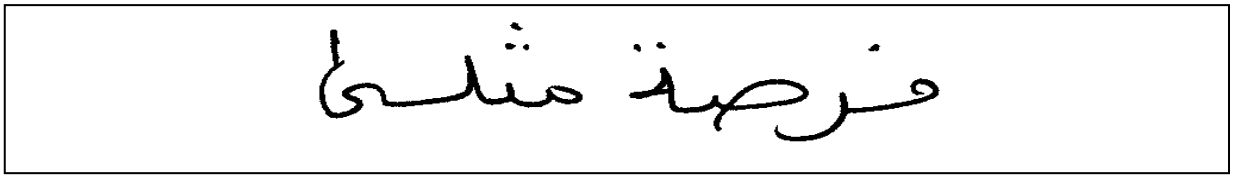

Figure 2: An example of handwritten Arabic text

Figure 3: An example of printed Latin text

Figure 4: An example of handwritten Latin text

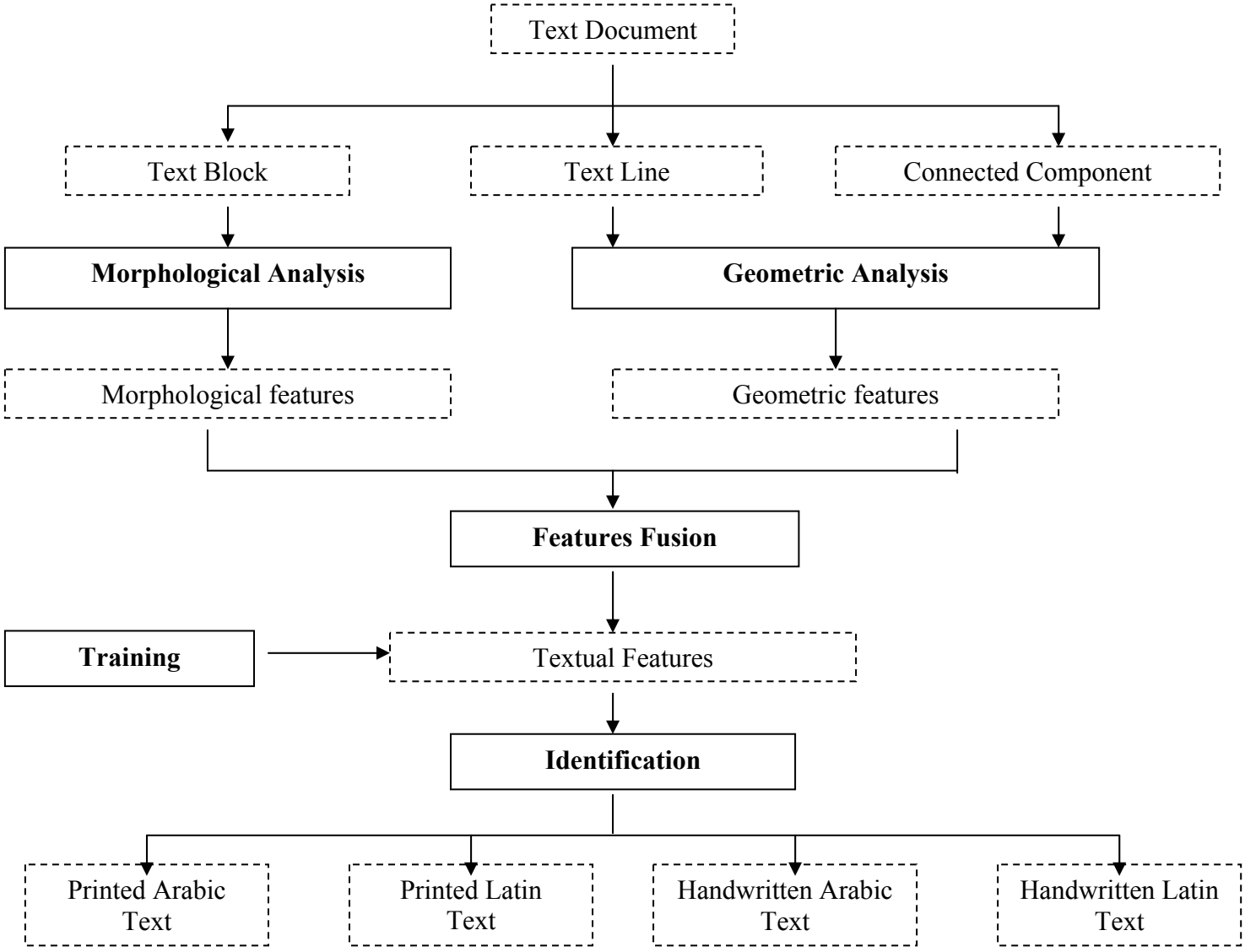

Figure 5: Global diagram of our identification approach

3. Experimental results

In [Kan00], we have presented results obtained on a data set of 400 blocks of text. The most of these blocks of text were of big size. This data set has been subdivided into 2 equal parts (200 for training and 200 for testing). We have used a KNN classifier without any optimization. The global rate of identification was 88.5 %.

To validate our approach on a little size of blocks of text, we constituted a data set of 335 blocks of text. Our data set was subdivided into 2 parts. The first part was used for training (222 blocks of text) and the other one for testing (113 blocks of text). We used two KNN classifiers (K = 5) : one without rejection (0 %) and one with maximum rejection (100 %). Therefore, we used many optimization of features extraction. The experimental results obtained are summarized in the following table:

KNN classifiers with	Decision	Rate
0 % rejection	Identification	92 %
	Confusion	0 %
	Rejection	8 %
100 % of rejection	Identification	69.5 %
	Confusion	27.5 %
	Rejection	3 %

We consider that our results are satisfactory compared to the data set on which we carried out our tests since it is of limited size and is not enough representative. Also, our results are encouraging owing to the fact that we treat at the same time the identification of the document script (Arabic and Latin) and of its nature (printed or handwritten). At present and to our knowledge, this problem is not mentioned yet in the literature. The results obtained within the framework of identification of script in the case of printed form are rather satisfactory. As an example, Spitz [Spi97] discriminates Latin of Asian with a rate of 100 % and Tan [Tan98] who identifies six scripts with a rate of 96.7 %. The results obtained by Hochberg et al [Hoc98] in the case of the handwritten is of about 88 % knowing that the authors treat Arabic and Latin plus four other scripts.

4. Conclusion and perspectives

In this paper, we proposed an approach to discriminate Arabic documents of Latin documents as well as their nature : printed or handwritten. Our approach is based on two steps: morphological on the text block level of text and another geometrical on the text line level of text and connected components. These two analyses are independent from the size of the analyzed blocks of text. Also, our global rates of correct identification is 88.5 % on data set of 400 big size of scanned blocks of text and is 92 % on data set of 335 little size of scanned blocks of text.

Currently, many improvements of our system are made in the following main axes:

- Enrich the features used in order to avoid the problems of confusion between printed text and handwritten text and between Arabic and handwritten Latin.
- The use of neural classifiers combined with KNN classifier [Rib99] should improve the results. The integration of an incremental classifier [Sto99] should help the automatic labeling of the images and so to easily increase the training data sets.
- Make tests on representative and large test data sets.

5. References

[Abd89] H.Y. Abdelazim, M.A. Hashish, , *Automatic reading of bilingual typewritten text*, Proc. Comp EURO'89 VLSI and Computer Peripherals, Hamburg, West Germany, pp. 2/140-144, May 1989.

[Hoc97] J Hochberg, P. Kelly, T. Thomas, L. Kerns, *Automatic Script Identification From Document Images Using Cluster-Based Templates*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 2, pp. 176-181, 1997.

[Hoc98] J Hochberg, K. Bowers, M. Cannon, P. Kelly, *Script and Language Identification for Handwritten Document Images,* International Journal on Document Analysis and Recognition, vol. 2, pp 45-52, 1999.

[Kan00] S. Kanoun, A. Ennaji, Y. Lecourtier, A. M. Alimi, *Une approche de identification Arabe / Latin, Imprimé / Manuscrit*, Proc CIFED'2000, Lyon, France, pp. 121-129, 2000.

[Pal97] U. Pal, B.B. Chaudhuri, *Automatic Separation of Words in Multi-lingual Multi-script Indian Documents*, Proc. Fourth Int'l Conf. Document Anal. Recog., Ulm-Germany, pp. 576-579, 1997.

[Lee96] D.S. Lee, C.R. Nohl, H.S. Baird, *Language Identification in Complex, Unoriented, and Degraded Document Images*, Proc. IAPR Workshop on Document Analysis Syst., pp. 76-98, Oct. 1996.

[Lee95] S.W. Lee, J.S. Kim, *Multi-lingual, Multi-font and Multi-size Large-set Character Recognition using Self-Organizing Neural Network*, Proc. ICDAR'99., Montreal, Canada, pp. 28-33, 1995.

[Rib99] A. Ribert, Y. Lecourtier, A. Ennaji, *Designing Efficient Distributed Neural Classifieurs : Application to Handwritten Digit Recognition*, Proc. ICDAR'99., Bangalore, India, pp. 265-268, 1999.

[Sto99] E. Stocker, Arnaud Ribert, Y. Lecourtier, *Self-organized Classification Problem Solving with Yprel Neural Networks*, Proc. ICDAR'99., Bangalore, India, pp. 265-268, 1999.

[Spi97] A.L. Spitz, *Determination of the the Script and Language Content of Document Images,* IEEE Trans. Pattern Analysis and Machine Intelligence vol. 19, no. 3, pp. 235-245, 1997

[Sue97] J. Ding, L. Lam, C.Y. Suen, *Classification of Oriental and European Scripts by Using Characteristic Features*, Proc. ICDAR'97, pp. 1023-1027.

[Tan98] T.N. Tan, *Rotation Invariant Texture Features and Their Use in Automatic Script Identification*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 7, pp. 751-756, 1998.

[Woo95] S. L Wood, X. Yao, K. Krishnamurthi, L. Dang, *Language Identification for Printed Text Independent of Segmentation*, Proc. IEEE ICIP'95, pp. 428-431, 1995.