Arabic and Latin script identification in printed and handwritten types Based on Steerable Pyramid Features

Mohamed Benjelil [1], Slim Kanoun [1], Rémy Mullot [2], Adel M. Alimi [1]
[1] *REGIM – ENIS, B.P. 1173, 3038, Sfax, Tunisia*
[2] *L3I, University of La Rochelle, Avenue Michel Crépeau, 17042. La Rochelle, France*

Abstract

Arabic and Latin script identification in printed and handwritten nature present several difficulties because the Arabic (printed or handwritten) and the handwritten Latin scripts are cursive scripts of nature. To avoid all possible confusions which can be generated, we propose in this paper an accurate and suitable designed system for script identification at word level which is based on steerable pyramid transform. The features extracted from pyramid sub bands serve to classify the scripts on only one script among the scripts to identify. The encouraging and promising results obtained are presented in this research paper.

1. Introduction

Current research in the field of document image analysis aims at conceiving and implementing automatic systems able to differentiate several scripts in order to select the recognition system appropriate to a given textual entity.

The produced documents each day in the whole world comprise different scripts, printed and handwritten types especially in the international administrative environments. Consequently, the automation of the script and the nature identification of the text blocks contained in document images became a need in any optical character recognition system.

The general framework of this paper is articulated around the script recognition systems in multilingual document images. The essential objective is to propose a strategy making it possible the reliable identification of Arabic and Latin scripts in printed and handwritten types.

In the sequel, we start initially with a synthesis of the existing systems for script and nature identification. We then propose a three decision levels strategy from Arabic and Latin texts identification in printed and handwritten types. Lastly, we achieve this paper by the experimental results obtained on a 1200 word images data set.

2. Related works

The script identification state of the art shows that the existing systems can be classified in to three main categories. Thus, we distinguish the systems based on the global analysis, the systems based on the local analysis and the systems based on the hybrid analysis.

The systems based on global analysis regard a text block as being only one entity and thus do not make recourse to other analyses from text line or word and connected component. Among these systems, Wood et al present in [16] a system based on the horizontal and vertical projection profiles analyses to discriminate Latin, Chinese and Arabic scripts for text blocks. In [2], [12] and [15], the authors propose a system based on the idea that the various scripts have different textures.

The systems based on local analysis are focused on the analysis of the intrinsic features of various scripts], [8],[20], [21], [22]. In [14], the proposed system uses the distribution of concavities to the top of the characters in the text lines to discriminate the Asian and the Latin scripts. This feature is also used in the same context in [9] with other features such as the distribution of concavities to the bottom, the heights distribution and alignments high and low of the characters. In [3], in addition to the concept of concavities distributions, the authors propose the horizontal projection profile analyses of the text lines, the heights distribution of the characters as well as the localization of the connected components within the same character (encased, at the top, at the bottom, on the right, on the left) to discriminate between European scripts and Oriental scripts. Within the same framework, several systems were presented in [4], [7], [11].

In the same category of systems, we distinguish also the systems based on the connected component models. These models are often obtained starting from training data sets. In this category, the system developed in [5] makes it possible to treat up to 13 different scripts by comparison between textual symbols extracted from text block to be identified and the connected component models. The same idea is

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.287

used in [6] to identify 6 different scripts in handwritten nature. In the same sense, other systems were proposed in [7] [17].

The systems based on hybrid analysis seek to develop script differentiation strategies exploiting all information available in the three principal levels of a textual entity script to identify: a text block, a text line or a word, and connected component. These strategies combine the global and local analyses. Within this framework, the suggested systems are based on the connected component analyses by adding another level of analysis which can be either the text line or the text block. The analyses used within the framework of this category of systems are similar to those used in the two categories of systems described above. Among these systems, we quote the systems described in [13].

3. Complexity of Arabic and Latin differentiation in printed and handwritten texts

The study of the systems presented in the precedent section shows that there are two strategy types to identify the script for any analysis categories (global, local, hybrid). The first is based on the features vector and a classifier while using training and test data sets. The second is based on the pre-established models or on the research of the existence or the absence of intrinsic features of each script. In the same framework, we notice that the first strategy type use only one decision level and only one features vector to differentiate at the same time the scripts. This one level could be enough if the scripts to be identified do not present some similarities.

In the same sense, the state of the art on the script identification shows that the existing systems treated either printed nature or handwritten nature. Within this framework, it is significant to note that the majority of systems are interested to printed nature rather than handwritten nature. Few systems treated handwritten nature [6]. This observation could be explained by the fact that the script identification is easier to printed nature than to handwritten nature. Thus, a printed text is uniform since it has some regularity in lines, words and letters of a given alphabet whereas a handwritten text is not uniform since it depends on the writing style of the writer. Also, we indicate that few systems are interested at the same time in printed and handwritten nature of the same script [1], [4], [8], [17]. The analysis of the results presented in [8], shows several confusions between the printed and the handwritten types for Arabic script and between this last script and the handwritten nature for Latin script. These confusions come mainly from the similarities between Arabic script for printed or handwritten types and handwritten nature from Latin script. This similarity is caused by the cursive nature of these last scripts. The other confusions found between the printed and the handwritten for Latin script are justified by the fact that many writers do not use the ligatures between the letters in their styles of writings in handwritten nature. To illustrate these similarities, we present an example of text block for Arabic (printed and handwritten) (Figure 1) and an example of Latin text block (printed and handwritten) (Figure 2).

To avoid all possible confusions which can be generated by the various similarities presented above, we propose in this paper a new texture descriptor based on Steerable Pyramids transform. Our motivation in using Steerable Pyramids relies not only on the fact that they have demonstrated discrimination properties for texture characterization [19], but also that unlike other image decomposition methods, the feature coefficients are less modified under the presence of image rotations, or even scales.

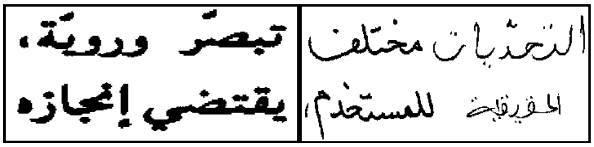

Figure 1.Left: Printed Arabic words, Right: Handwritten Arabic words

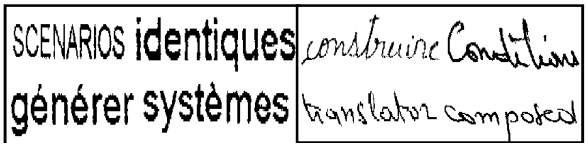

Figure 2.Left: Printed Latin words, Right: Handwritten Latin words

4. Steerable pyramid (S.P)

The Steerable Pyramid [18], is a linear multi-scale, multi-orientation image decomposition, that provides a useful front-end for image-processing and computer vision applications. The S.P can capture the variation of a texture in both intensity and orientation.

The synoptic diagram for the decomposition (both analysis and synthesis) is shown in (Figure 3). Initially, the image is separated into low and high pass sub bands, using filters L_0 and H_0. The low pass sub band is then divided into a set of oriented band pass sub bands and a lower pass sub band. This lower pass sub band is sub sampled by a factor of 2 in the X and Y directions. The recursive (pyramid) construction of a pyramid is achieved by inserting a copy of the shaded portion of the diagram at the location of the solid

circle. The basic functions of the steerable pyramid are directional derivative operators that come in different sizes and orientations.

The necessary conditions for a filter basis to be steerable, is the ability to synthesize a filter of any orientation from a linear combination of filters at fixed orientations (Figure. 4).

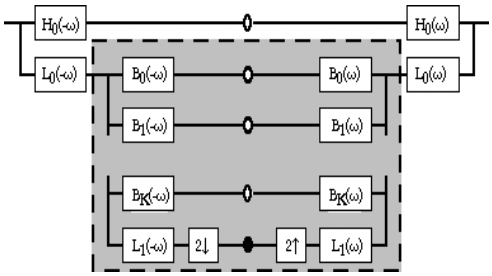

Figure 3.First level of steerable pyramid decomposition [18]

The simplest example of this is oriented first derivative of Gaussian filters, at 0° and 90°:

$\alpha_1 = 0°, \alpha_2 = 90°$

The steering equation:

$$G_1^{\alpha}(x,y) = \cos(\alpha)G_1^{0°}(x,y) + \sin(\alpha)G_1^{90°}(x,y)$$

We can synthesize a filter at any orientation by linear combination of filters $G_1^{0°}$ and $G_1^{90°}$. We can synthesize an image at any orientation by linear combination of the convolution of that image with the filters $G_1^{0°}$ and $G_1^{90°}$:

For $R_1^{0°} = G_1^{0°} * I$ and $R_1^{90°} = G_1^{90°} * I$

The resulting image is

$$R_1^{\alpha}(x,y) = \cos(\alpha)G_1^{0°}(x,y) + \sin(\alpha)G_1^{90°}(x,y)$$

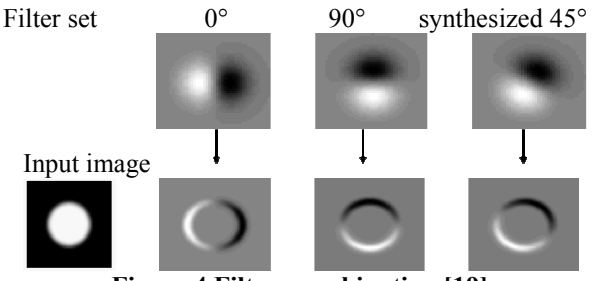

Figure 4.Filters combination [19]

We tested the S.P on 300 printed Arabic text bloc, 300 printed Latin text bloc, 300 handwritten Arabic text bloc and 300 handwritten Latin text bloc. This particular steerable pyramid contains 4 orientation sub bands, at 2 scales each. For each image sub bands, we calculated the variance, the mean, the homogeneity and energy. The obtained results encourage us to use it script identification. The scatter plots, (Figure .5), show clearly how the features measurements differs between scripts. We use just the two columns containing the mean and standard deviation measurements.

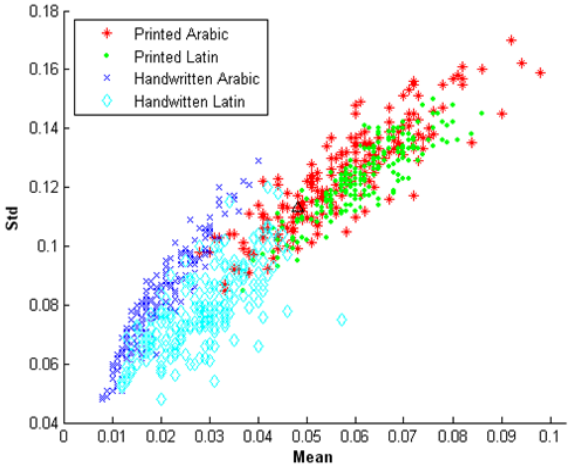

Figure 5.Training data set scatter plot

4. S.P based script identification strategy

We consider text and non text as regions with different textures. Since the distinguishing characteristics of text are frequency information, orientation, approximately with the same size and line thickness, located at a regular distance from each other, we can use them to characterize text regions with steerable pyramid decomposition. (Figure .6) shows the synoptic diagram of proposed system.

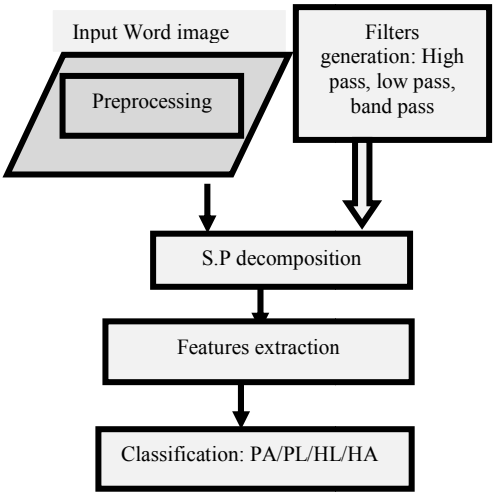

Figure 6.Synoptic diagram of proposed system

S.P decomposition: We used a steerable pyramid with 4 orientation sub bands, at 2 scales. The number of orientations may be adjusted by changing the

derivative order (for example, the first derivatives yield two orientations). In figure 7 and 8, we show a sample decomposition of two text blocs

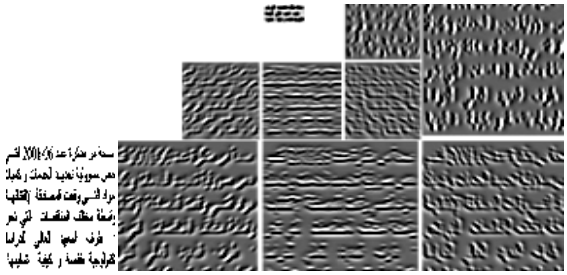

Figure 7.Steerable pyramid decomposition with 2 levels and 4 orientations of printed Arabic text block

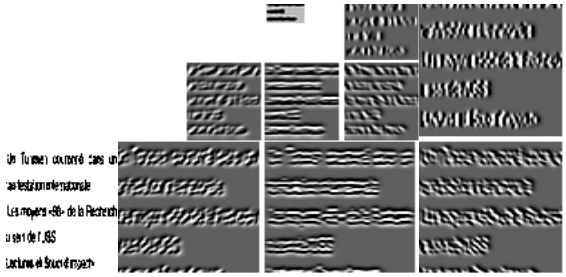

Figure 8.Steerable pyramid decomposition with 2 levels and 4 orientations of printed Latin text block

Features extraction: The feature vector (with dimension 90 to represent 18 sub bands) was constructed based on the computed mean μ_{mn}, the standard deviation σ_{mn}, the kurtosis K_{mn} of the magnitude of the transformed word image and the energy E_{mn}, the homogeneity H_{mn} and the correlation C_{mn} calculated from gray-level co-occurence matrix applied to the same transformed word image. This feature vector is defined as:

$$FV = [\mu_{11}\ \sigma_{11}\ E_{11}\ H_{11}\ C_{11}\ \ldots\ \mu_{18}\ \sigma_{18}\ E_{18}\ H_{18}\ C_{18}]$$

Where,

- Mean
$$\mu = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N} I_i(x,y)}{M \times N}$$
- Standard deviation
$$\sigma^2 = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N} [I_i(x,y) - \mu]^2}{M \times N}$$
- Kurtosis
$$k = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N} [I_i(x,y) - \mu]^4}{M \times N \times \sigma^4} - 3$$
- Energy
$$E = \sum_{i,j}^{n} p(i,j)^2$$
- Homogeneity
$$H = \sum_{i,j} \frac{p(i,j)}{1 + |i-j|}$$
- Correlation
$$C = \sum_{i,j} \frac{(i-\mu i)(j-\mu j)\hat{P}(i,j)}{\sigma i \sigma j}$$

5 Results and discussion

To try out the proposed strategy, we constituted a data set of 800 word images including 200 of each class: printed Latin, printed Arabic, handwritten Arabic and handwritten Latin. We then subdivided this data set in two data sets: a data set for the training and a data set for the test containing each one 100 word images of each class.

For scripts (Arabic and Latin) and types (printed and handwritten) identification, we used the K nearest neighbors classifier with K=3, 5 and 7. We found that the best identification rates are obtained with K=5.

The S.P parameters tested are sp0filter, sp3filter, sp5filter with respectively 2, 4, 6 orientations and 1, 2, 3 and 4 levels. We found that the best identification rates are obtained by sp3filter with 4 orientations and 2 levels.

Table 1 synthesizes the correct identification rates obtained our proposed system. In this table, we used CI, C, PA, PL, HA and HL abbreviations respectively to represent correct identification rate, Confusion rate, Printed Arabic, Printed Latin, Handwritten Arabic and Latin Handwritten**.** The overall correct identification rate obtained is about 97.5 %.

Table 1.Correct identification rates of proposed strategy.

Script and Nature	%CI	%C	Confusion matrix			
			PL	PA	HA	HL
PL	99%	1%	99	1	0	0
PA	98%	2%		98	1	1
HA	97%	2%			97	3
HL	96%	4%			4	96
Moy	97.5%	2.5%				

The analysis of the results presented in table 1 shows that our strategy proposed in this paper could

identified in a reliable way Latin printed with a correct identification rate about 99 %. On the other hand, there are some confusion between the handwritten Arabic and the handwritten Latin because their cursive nature.

8. Conclusion and future work

The work developed in this paper aims at setting up a system of differentiation between the Arabic and the Latin script in printed and handwritten types. Thus, we begin with a study from the existing systems of script differentiation. Within this framework, we showed that the majority of systems are interested in printed nature. Few systems treated handwritten nature. We then proposed a strategy which is based on steerable pyramid transform.

Currently, the improvements of the proposed strategy are to combine the local and global text block analyses and especially to solve the confusion problems which still exist between Handwritten Arabic and Handwritten Latin scripts.

References

[1] Bennasri, A., Zahour, A., Taconet, B., 2000. Arabic Script Preprocessing and Application to Postal Addresses. Proc. ACIDCA'2000, 74-79.

[2] Busch, A., Boles, W. W., Sridharan, S., 2005. Texture for Script Identification. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 27, no. 11, 1720-1732.

[3] Ding, J., Lam, L., Suen, C.Y., 1997. Classification of Oriental and European Scripts by Using Characteristic Features. Proc. International Conference on Document Analysis and Recognition, 1023-1027.

[4] Fan, K., Wang, L., Tu, Y., 1998. Classification of machine-printed and handwritten texts using character block layout variance. International Journal of Pattern Recognition, vol. 31, no. 9, 1275-1284.

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

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

[7] Jaeger, S., Ma, H., Doermann, D., 2005. Identifying Script on Word-Level with Informational Confidence. Proc. International Conference on Document Analysis and Recognition, 416-420.

[8] S. Kanoun, A. Ennaji, A. Alimi, Y. Lecourtier "Script and Nature Differentiation For and Latin Text Images", *8th IAPR - International Workshop on Frontiers in Handwriting Recognition : IWFHR'2002*, pp. 309 - 313, 6-8 Août, 2002, Niagara-on-the-Lake, Ontario, Canada.

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

[10] Liu, Y. H., Lin, C. C., Chang, F., 2005. Language Identification of Character Images Using Machine Learning Techniques. Proc. International Conference on Document Analysis and Recognition, 630 – 634.

[11] Pal, U., Sinha, S., Chaudhuri, B. B., 2003. Multi-Script Line identification from Indian Documents. Proc. International Conference on Document Analysis and Recognition, 880 – 884.

[12] Pan, W. M., Suen, C. Y., Bui, T. D., 2005. Script Identification Using Steerable Gabor Filters. Proc. International Conference on Document Analysis and Recognition, 883-887.

[13] Patil, S. B., Subbareddy, N. V., 2002. Neural network based system for script identification in Indian documents. Sadhana, vol. 27, Part 1, 83–97.

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

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

[16] Wood, S. L., Yao, X., Krishnamurthi, K., Dang, L., 1995. Language identification for Printed Text Independent of Segmentation. Proc IEEE International Conference on Image Processing, 428-431.

[17] Zhou, L., Lu, Y., Tan, C. L., 2006. Bangla/English script identification based on analysis of connected component profiles. Proc. 7th IAPR workshop on Document Analysis Systems.

[18] Simoncelli, E.P., Freeman, W.T., 1995. The steerable pyramid: A flexible architecture for multi-scale derivative computation, In: Proc. IEEE Second Internat. Conf. on Image Process. Washington, DC, pp. 444–447.

[19] W. T. Freeman, E. H. Adelson, «The design and use of steerable filters», IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 13, no. 9, pp. 891-906,September,1991.

[20] M. Benjelaiel, S.Knoun, A. Alimi, R. Mullot, "Three decision levels strategy for Arabic and Latin texts differentiation in printed and handwritten natures", *9th IAPR - International Conference on Document Analysis and Recognition: ICDAR'2007*, pp. 1103 - 1107, Volume 2, 23 – 26 September, 2007, Curitiba, Paraná, Brazil.

[21] S. Kanoun, I. Moalla, A. Ennaji, A. Alimi "Script Identification for Arabic and Latin, Printed and Handwritten Documents", *4th IAPR - International Workshop on Document Analysis Systems : DAS'2000*, pp. 159-165, 10 -13 Décembre 2000, Rio de Janeiro, Brazil.

[22] S. Kanoun, A. Ennaji, A. Alimi, Y. Lecourtier."Une approche de discrimination Arabe / Latin, Imprimé / Manuscrit", *2ème Colloque International Francophone sur l'Ecrit et le Document : CIFED'2000*, pp. 121-129, 3 – 5 Juillet 2000, Lyon, France.