Language and script identification based on Steerable Pyramid Features

Mohamed Benjelil
REGIM – ENIS, B.P. 1173, 3038, Sfax, Tunisia
L3I, Univ. of La Rochelle, France
mohamed.benjlaiel@ieee.org

Rémy Mullot
L3I, University of La Rochelle, Avenue Michel Crépeau, 17042. La Rochelle, France
remy.mullot@univ-lr.fr

Adel M. Alimi
REGIM – ENIS, B.P. 1173, 3038, Sfax, Tunisia
adel.alimi@ieee.org

Abstract

Arabic and Latin language and script identification in machine printed and handwritten types present several difficulties because the Arabic (machine 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 language and 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, machine printed and handwritten types especially in the international administrative environments. Consequently, the automation of the script and the type 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 machine printed and handwritten types.

In the sequel, we start initially with a synthesis of the existing systems for script and type identification. We then propose a three decision levels strategy from Arabic and Latin texts identification in machine 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,

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.226

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 used in [6] to identify 6 different scripts in handwritten type. 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 machine 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 machine printed type or handwritten type. Within this framework, it is significant to note that the majority of systems are interested to machine printed type rather than handwritten type. Few systems treated handwritten type [6]. This observation could be explained by the fact that the script identification is easier to machine printed type than to handwritten type. Thus, a machine 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 machine printed and handwritten type of the same script [1], [4], [8], [17]. The analysis of the results presented in [8], shows several confusions between the machine printed and the handwritten types for Arabic script and between this last script and the handwritten type for Latin script. These confusions come mainly from the similarities between Arabic script for machine printed or handwritten types and handwritten type from Latin script. This similarity is caused by the cursive nature of these last scripts. The other confusions found between the machine 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 type. To illustrate these similarities, we present an example of words images for Arabic (machine printed and handwritten) (Figure 1) and an example of Latin words images (machine printed and handwritten) (Figure 2).

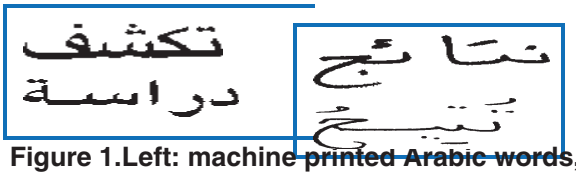

Figure 1.Left: machine printed Arabic words, Right: Handwritten Arabic words

Figure 2.Left: machine printed Latin words, Right: Handwritten Latin words

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.

4. Steerable pyramid (S.P)

Steerable Pyramid Decomposition [18], is a linear multi-orientation, multi-resolution image decomposition method, by which an image is subdivided into a collection of sub-bands localized at different scales and orientations.

The synoptic diagram for a first level image decomposition using Steerable Pyramid is shown in Figure 5. Using a high-pass and low-pass filter (*H0, L0*) the input image is initially decomposed into two sub-bands: a high-pass, and a low-pass sub-band, respectively. Further, the low-pass sub-band is decomposed into *K*-oriented band-pass portions *B*0 to *BK*−1, and into a low-pass sub-band *L*1. The decomposition is done recursively by sub-sampling the low-pass sub-band by a factor of 2 along the rows and columns. Each recursive step captures different directional information such as variation of a texture in both intensity and orientation at a given scale.

The basic functions of the steerable pyramid are directional derivative operators that come in different sizes and orientations. The simplest example of this is the oriented first derivative of Gaussian. Consider the two-dimensional circularly symmetric function G written in Cartesian coordinates x and y:

$$G(x, y) = e^{-(x2+y2} \quad (1)$$

The scaling and normalization constants have been set to 1 for convenience. The directional derivative operator is steerable as is well-known [11], Let us write the *n*th derivative of a Gaussian in the *x* direction as G_n. Let $(\ldots)^{\theta}$ represent the rotation operator such that for any function *f(x, y)* , $f^{\theta}(x, y)$ is *f(x, y)* rotated through an angle θ about the origin. The first *x* derivative of a Gaussian $G_1^{0^\circ}$ is

$$G_1^{0^\circ} = \frac{\partial}{\partial x} e^{-(x^2+y^2)} = -2xe^{-(x^2+y^2)} \quad (2)$$

The same function, rotated 90°, is

$$G_1^{90^\circ} = \frac{\partial}{\partial y} e^{-(x^2+y^2)} = -2ye^{-(x^2+y^2)} \quad (3)$$

These functions are shown in Figure 3. It is straightforward to show that a G_1 filter at an arbitrary orientation θ can be synthesized by taking a linear combination of $G_1^{0^\circ}$ and $G_1^{90^\circ}$.

$$G_1^{\theta} = \cos(\theta)\, G_1^{0^\circ} + \sin(\theta)\, G_1^{90^\circ} \quad (4)$$

Since $G_1^{0^\circ}$ and $G_1^{90^\circ}$ span the set of G_1^{θ} , we call them basis filters for G_1^{θ}. The $\cos(\theta)$ and $\sin(\theta)$ terms are the corresponding interpolation functions for those basis filters. Because convolution is linear operation, we can synthesize an image filtered at any arbitrary orientation by taking linear combinations of the images filtered with $G_1^{0^\circ}$ and $G_1^{90^\circ}$. Letting * represent convolution and I the input image, for $R_1^{0^\circ} = G_1^{0^\circ} * I$ and $R_1^{90^\circ} = G_1^{90^\circ} * I$, the resulting image is

$$R_1^{\theta} = \cos(\theta)\, R_1^{0^\circ} + \sin(\theta)\, R_1^{90^\circ} \quad (5)$$

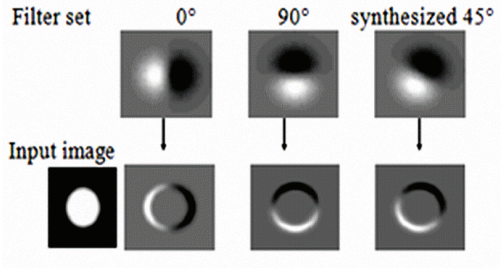

Figure 3.Filters combination [19]

We tested the S.P on 300 machine printed Arabic text bloc, 300 machine printed Latin text bloc, 300 handwritten Arabic text blocs and 300 handwritten Latin text blocs. 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 4), show clearly how the features measurement differs between scripts. We use just the two columns containing the mean and standard deviation measurements for sub band 2 and scale1.

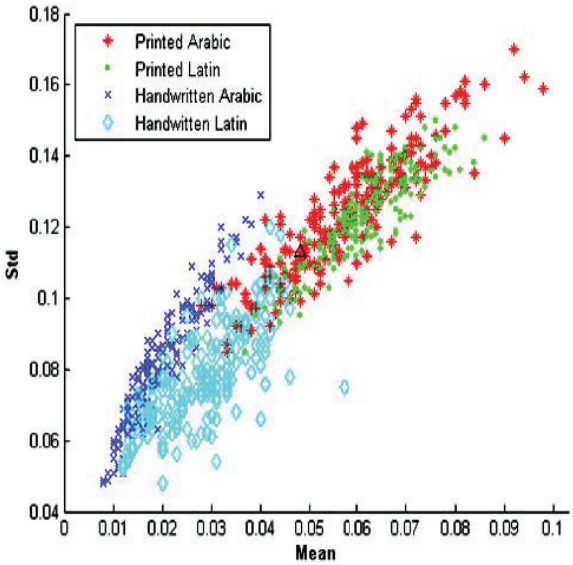

Figure 4.Training data set scatter plot

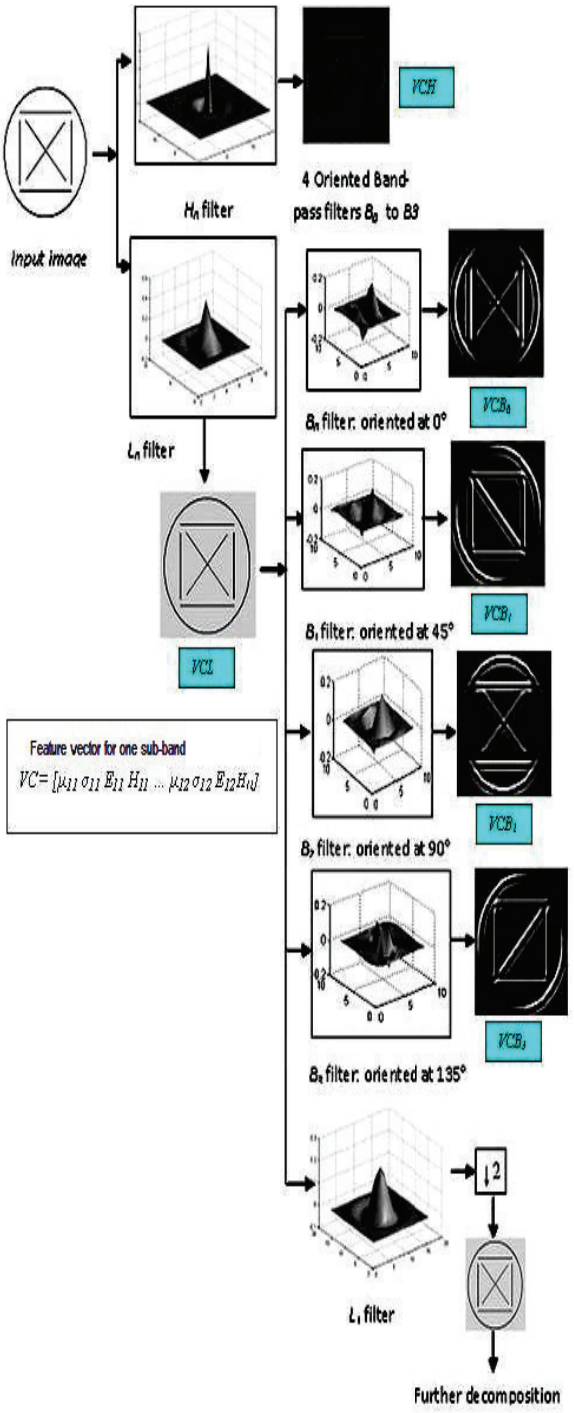

Figure 5.First level of steerable pyramid decomposition [18]

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

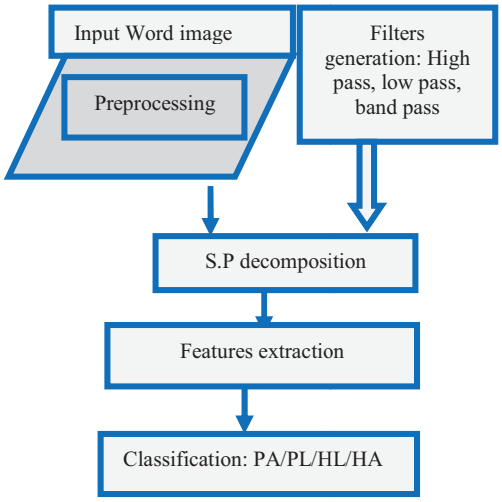

Figure 6. Synoptic diagram of proposed system

S.P decomposition:

Figure 7.S.P decomposition with 2 levels and 4 orientations of handwritten Arabic word

Figure 8.S.P decomposition with 2 levels and 4 orientations of handwritten Latin word

Features extraction:

(with dimension 48 to represent 8 sub bands)
The feature vector was constructed using the computed mean μmn, the standard deviation σmn, 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-occurrence matrix applied to the same transformed word image. Supposing using *M* scales and N orientations. This feature vector is defined as:

$F = [\mu_{00}\ \sigma_{00}\ K_{00} E_{00}\ H_{00}\ C_{00}\ \ldots\ \mu_{mn}\sigma_{mn}\ K_{mn}\ E_{mn}\ H_{mn}\ C_{mn}]$, m=0,1,…,M, n=0,1,…,N.

Where,

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

P(i,j) is the probability density that the first pixel has intensity value i and the second j.

6. Results and discussion

To try out the proposed strategy, we constituted a data set of 800 word images including 200 of each class: machine printed Latin, machine 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 (machine 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, machine printed Arabic, machine 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 type	%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 machine 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.

7. 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 machine 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 machine printed type. Few systems treated handwritten type. 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.