Proposition to distinguish Machine-Printed from Handwritten Arabic and Latin Words

Asma Saïdani, Afef Kacem Echi, Abdel Belaïd

To cite this version:

Asma Saïdani, Afef Kacem Echi, Abdel Belaïd. Proposition to distinguish Machine-Printed from Handwritten Arabic and Latin Words. Information and Communication Technologies Innovations and Applications (ICTIA), Information and Communication Technologies Innovations and Applications (ICTIA), Mar 2014, Sousse, Tunisia. hal-01112678

HAL Id: hal-01112678

https://hal.archives-ouvertes.fr/hal-01112678

Submitted on 3 Feb 2015

Proposition to distinguish Machine-Printed from Handwritten Arabic and Latin Words

Asma Saïdani and Afef Kacem Echi
University of Tunis, LaTICE-ENSIT
Tunis, Tunisia
saidaniasma@yahoo.fr
afef.kacem@esstt.rnu.tn

Abdel Belaïd
University of Lorraine, LORIA
Nancy, France
Abdel.belaid@loria.fr

***Abstract*—In this work, we gathered some contributions to identify script and its nature. We successfully employed many features to distinguish between handwritten and machine-printed Arabic and Latin scripts at word level. Some of them are previously used in the literature, and the others are here proposed. The new proposed structural features are intrinsic to Arabic and Latin scripts. The performance of all extracted features is studied towards this paper. We also compared the performance of three classifiers: *Bayes* (*AODEsr*), *k-Nearest Neighbor* (*k-NN*) and *Decision Tree* (*J48*), used to identify the script at word level. These classifiers have been chosen enough different to test the feature contributions. We carried experiments using standard databases. Obtained results demonstrate used feature capability to capture differences between scripts. Using a set of 58 selected features and a *Bayes*-based classifier, we achieved an average identification rate equals to 98.72%, which considered a very satisfactory rate compared to some related works.**

Keywords—* *Script and nature identification; machine-printed/handwritten word, Arabic/Latin script; Feature extraction; Classification;

I. Introduction

The term script identification refers to the task of identifying the language a given document is written in. It plays a major role in several applications. It is mostly used as an important preprocessing step in the design of an optical character recognition system. Some works have been done on script identification in past years. These woks depend on many types of features extracted from document images at block, text-line or word level. Block level script identification identifies the script of the given document in a block and concerns documents written in different languages. In text-line based script identification, a document image can contain more than one script and the script is written occasionally to highlight one sentence. Word level script identification allows the document to contain more than one script and the words are scattered throughout the document whenever it is necessary. Notice that most of the existing systems work on block level script identification. As they are based on the overall visual appearance of the text-block, they are generally incapable of tackling the variations in the writing style, character style and size, spacing between lines or words, etc. When the classification is performed by words and not by text-line or text-block, it will be possible to analyze more cases with scripts more or less long, written in the form of words or lines. But this requires finest analysis of each word [14].

This paper aims identifying the script (Arabic or Latin) and its nature (printed-machine or handwritten), at the word level. We extracted many features, tested and evaluated them under the same experimental conditions. The objective is to contribute to the field of script and nature identification through better selection and combination of features, used in the literature, with those here proposed. The remainder of the paper is organized as follows. In section II, we describe Arabic and Latin script characteristics. Then we present, in section III, some related works. Next, we emphasize on feature extraction and selection and discuss experimental results in sections IV and V. Finally, we conclude in section VI and propose some future directions.

II. Arabic and Latin Script Characteristics

Arabic is different from Latin with respect to a number of aspects. Arabic alphabet is written from right to left, in a cursive style, and includes 28 letters without considering the variation of their shapes according to the position, the voyellation elements and the phonetic context. There are no distinct upper and lower case letter forms. Many letters look similar but are distinguished from one another by dots above or below their central part. These dots are an integral part of a letter, since they distinguish between letters that represent different sounds. For example, the Arabic letters ب and ت have the same basic shape, but ب has one dot below and ت has two dots above. In fact, Arabic writing is very rich in diacritic marks (e.g. dots, hamza, etc.). Diacritic points can be located above or below the letter, but never both simultaneously. A Latin text is lower in diacritic compared to an Arabic text. There are only the two letters 'i' and 'j' which have only one diacritic point above. There are no diacritic at the bottom in a Latin text. Unlike cursive writing based on the Latin alphabet, the standard Arabic style is to have a substantially different shape depending on whether it will be connecting with a preceding and/or a succeeding letter, thus all primary letters have conditional forms, depending on whether they are at the beginning, middle or end of a word, so they may exhibit four distinct forms (initial, medial, final or isolated). However, six letters (ا د ذ ر ز و) have only an isolated or final form. Some letters look almost the same in all four forms, while others show considerable variation. Both printed and written Arabic are cursive, with most of the letters within a word directly connected to the adjacent letters. In addition, some letter combinations are written as ligatures (special shapes such the Arabic letter ل and ا combination as لا and ـلا). There is no distinction between printing and cursive, as there is the case in Latin. Arabic word is a sequence of connected components called PAWs (Piece of Arabic Word). Each PAW is a sequence of completely cursive letters. PAWs are separated by small blanks and not necessarily composed of the same number of letters. The foregoing features relate to the Arabic writing whether printed or handwritten. In case of manuscript, others specificities are involved. We can quote fusion of diacritical points: Two or three diacritical points can easily be agglomerated in two or even one diacritical point. Also, in one word, two consecutive PAWs may overlap. For more details about Arabic/Latin script properties, see previous works in [1] and [2].

III. Related Works

Table I summarizes the surveyed methods. They mainly concern Arabic/Latin and handwritten/machine-printed scripts identification.

TABLE I. Identification Method Summarization

Ref.	***Script***	***Nature***	***Level***	***Id. Rate***
[3]	A/L	P/H(400B)	B/T/C	88.5%
[4]	F	P/H (32006W)	W	97.1%
[5]	L	P/H (Public databases)	W	> 80%
[6]	G/E	H (1200B)	C	95%
[7]	A/L	P/H (400B)	B	95%
[8]	L	P/H (50D)	T	98.2%
[10]	A/E	P (1976T,8320W)	T/W	99.7%(T), 96.8%(W)
[11]	A/L	P/H (800D)	B	84.75%
[12]	A/L	P/H (800W)	W	97.5%
[13]	A/L	P(learning: 3383W, test: 846W)	W	94.32%
[24]	A/L	P/H(learning: 400D, test: 200D)	D	82%(A), 92%(L)
[2]	A/R	P/H(1320W)	W	98.4%

A: Arabic, L: Latin, E: English, F: Farsi_Arabic, G: Bangla, R: French, D:Document, B: Block, T: Text-line, W: Word, C: Connected component, P: Printed, H: Handwritten.

IV. Feature Extraction

Several feature sets, used in the literature or proposed here, are used to illustrate their properties and performances.

A. Features Proposed in the Literature :

- Vertical projection variance of the word: Due to overlaps between handwritten words, projection profiles has smoother valleys and peaks resulting in smaller variance compared to machine-printed words [11]. The variance of the vertical co-ordinates of the vertical projection profile is calculated as a measure of homogeneity of the projection profile. In our point of view, this feature can be used to separate handwritten from machine-printed words either they are written in Arabic or Latin (see Fig. 1).

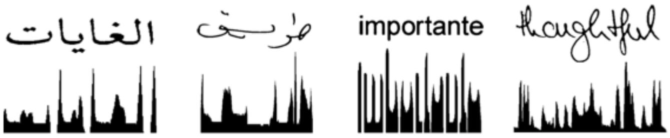

Fig. 1. Vertical projection variance of the word.

- Connected component width, height, aspect ratio, area and density: As underlined by [7], the sizes of connected components inside a machine-printed word are more consistent, leading to smaller width and height variances. Each of these features is generalized in terms of mean and standard deviation. According to us, these features can only discriminate between handwritten and machine-printed Latin words but not between handwritten and machine-printed Arabic words. In fact, in Arabic there is no distinction between printing and cursive. Moreover, the same Arabic word does not have a fixed length on account of the elongations.

- Separator length between two successive connected components: In printed Latin, connected components are separated by regular separators as noted by [7] and shown in Fig. 2. This feature is generalized in term of mean and standard deviation. In our opinion, this feature can just identify the nature (handwritten/machine-printed) of Latin words but not of Arabic words because of PAW overlapping.

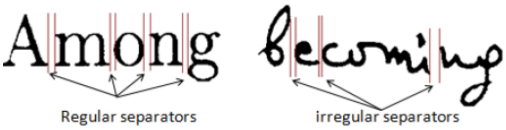

Fig. 2. Separator length between two successive connected components.

- Connected component profiles analysis: Both Handwritten and machine-printed Arabic is cursive. It is also the case of handwritten Latin but not for machine-printed Latin. As done in [6], we extracted the bottommost profile of the connected components (after elimination of diacritic points) which is the lowest pixels of vertical columns of the components. To obtain the bottommost profile, each vertical column of a particular connected component is scanned from bottom until it reaches a black pixel P_i. Thus, for a component of width N, we get N such pixels. For examples of bottommost profiles, see Fig. 3. To measure the discontinuity of bottommost contour line of the component, we compute difference d_i of two adjacent pixels P_i and P_{i+1} of the components. Note that Arabic script has lower discontinuity since it is straighter and flat and does not have high links between letters like in handwritten Latin script (especially in case of o and r letters). In fact, the high links increase the differences between coordinates of lower profile pixels of word.

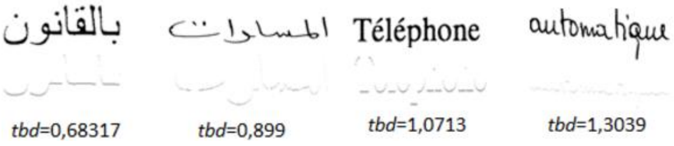

Fig. 3. Connected component profiles analysis.

- Loop aspect ratio: Printed script is a succession of connected components comprising loops whose surfaces are regulars contrary to the handwritten Arabic and Latin which depend on the writing style [3] (see Fig. 4). The loop aspect ratio is considered as feature and generalized in terms of mean and standard deviation.

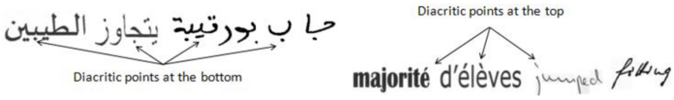

Fig. 4. Loop aspect ratio.

- Pixels distribution: This feature is especially used to discriminate handwritten from printed Latin in [8]. The bounding box is divided in two by a horizontal line as indicated by the red line (see Fig. 5). The bounding box height is decreased by 10 pixels as shown by the green lines. Then the density of the upper part and of the lower part is calculated. We retain the difference between these densities as feature.

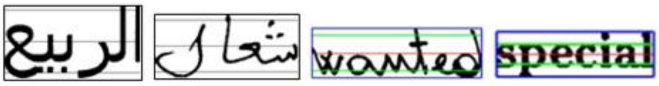

Fig. 5. Bounding box horizontal division.

Baseline profile features: Most pixels of machine-printed words are located on the baseline. In the machine-printed words, position of ascender and descender is determined by the baseline. However, in handwritten case words are not usually written in a single baseline and position of ascender and descender varies according to the writer's style. Difference between handwritten and machine-printed words is shown in the baseline profile features [4]. The baseline is estimated as the peak of horizontal histogram of the word image (see Fig. 6). The following features are extracted from the baseline profile: baseline position, sub-baseline number *n* (a sub-baseline represents pixels of the word image that lays on the baseline as shown in Fig. 7), distance of highest scan line from the baseline (d_1), distance of lowest scan line from the baseline (d_2) and the number of pixels on the baseline (p). Mean, variance of sub-baselines, and ratio of sub-baseline to their variances are also taken as features.

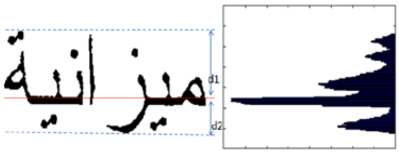

Fig. 6. Baseline profile features.

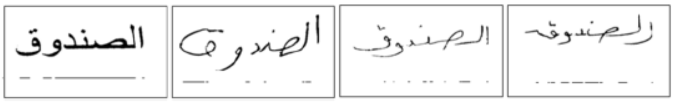

Fig. 7. Sub-lines of an Arabic word.

- Run-length histogram: In [9], authors proposed to extract features from Run-length histogram for machine-printed/handwritten Chinese character classification. Note that these features can be used to underline the difference between the stroke length of machine-printed and handwritten words. We extracted black pixel run-lengths in three directions, including horizontal, vertical and diagonal. We then calculated three histograms of run-lengths for these directions. To get scale-invariant features, we normalized the histograms. To get the final features, the histogram is then divided into five bins with equal width and five rectangular-shaped weight windows are used. Thus, we extracted five features in each direction, leading to 15 features. Fig. 8 presents the black vertical run-length of some words. We noted that run length values are high for handwritten words.

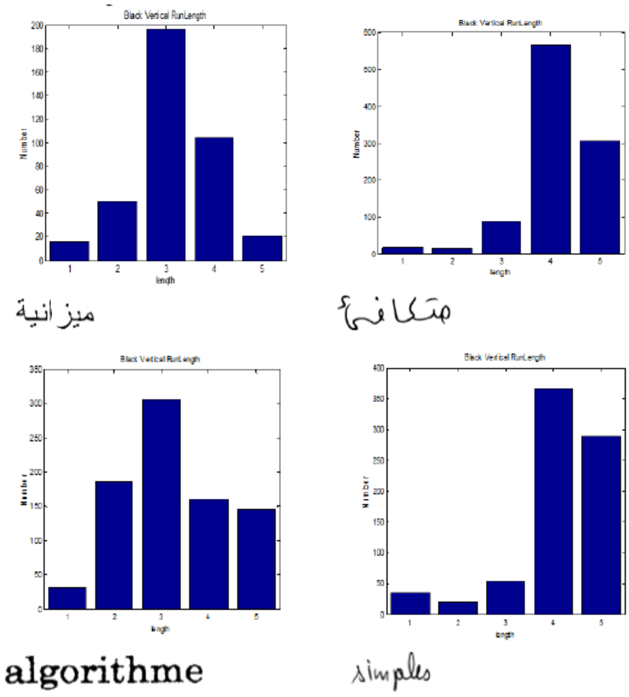

Fig. 8. BlackVertical Run length histogram.

- Crossing Count Histogram: Crossing count is the number of transitions from 0 to 1 along a hypothetical horizontal or vertical line over the word image. Crossing count features are already used by [4] for Latin handwritten and machine-printed Farsi_Arabic words discrimination. These features are used to measure stroke complexity. For each horizontal and vertical scan line, we computed the crossing count. To have the final features from the histograms, the same technique, used to extract the run-length features, is exploited

- Bi-level Co-occurrence: As defined in [15], a co-occurrence count is the number of times a given pair of pixels occurs at a fixed distance and orientation. For binary images, the possible co-occurrence pairs are white-white, black-white, white-black and black-black. As the black-black pairs carry most of the information than the other co-occurrence pairs, we only considered them to extract related features. We used horizontal, vertical, major and minor diagonal orientations and 2 pixels distance level for the classification.

- Upper lower Profile: In [8], authors tried to discriminate machine-printed from handwritten Latin text, using simple structural characteristics based on the fact that the height of printed characters is more or less stable within text-line. On the other hand, the distribution of the height of handwritten characters is quite diverse. These remarks stand also for the height of the main body of the character as well as the height of both ascenders and descenders. Thus the ratio of ascender height to main body's height and the ratio of descender's height to main body's height would be stable in printed text and variable in handwriting. To characterize a word, based on its upper lower profile, we extracted the following features: the ratio of ascender zone to the main body zone, the ratio of descender zone to the main body zone and the ratio of the area to the maximum value of the horizontal histogram of the upper-lower profile. Fig. 9 gives an example of these features computing on machine-printed Arabic word. Notice that connected components of diacritic points are not considered in the analysis of the upper lower profile.

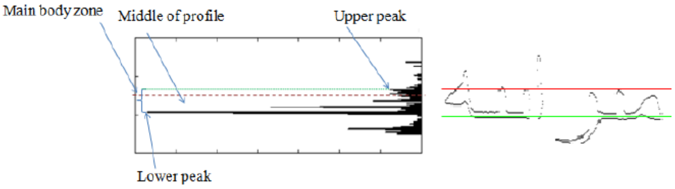

Fig. 9. Upper Lower profile.

- Word Physical sizes: [4] noted that the sizes of machine-printed words are more consistent than those of handwriting on the same form. Thus, features related to the physical sizes of the word block such as density of black pixels, width, height, aspect ratio and area are considered.

- Overlapping areas: Unlike machine-printed words, for a handwritten word, the bounding boxes of the connected components tend to overlap with each other [4]. The overlapping area, normalized by the total area of the block is calculated as feature (see Fig. 10).

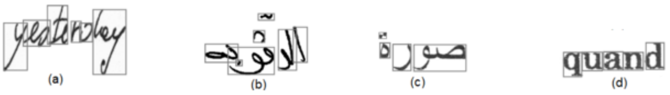

Fig. 10. Overlapping areas: (a) and (b) handwritten words, (c) and (d) machine-printed words.

- Moments: Simple properties of the word image can be found via image moments. They include area, its centroid and information about its orientation. So, we considered both central and *Hu* moments.

- Steerable Pyramid Transform: Steerable pyramid decomposition 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 (see Fig. 11). Features extracted from pyramid sub-bands served, in [12] to classify the scripts on only one script among the scripts to identify.

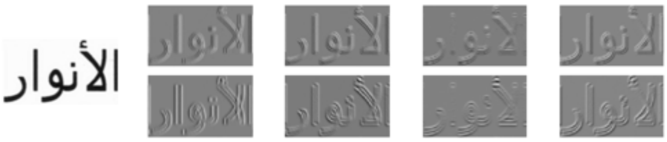

Fig. 11. Decomposition with 2 levels and 4 orientations of a printed Arabic word.

- Gabor filters: Gabor filter is a linear filter used for edge detection. Frequency and orientation representations of Gabor filters are similar to those of the human visual system, and they have been found to be particularly appropriate for texture representation and discrimination. In [16], Gabor filters are applied and 16 channels of features are extracted to identify the script (English or Chinese) of machine-printed words in scanned document images. In [24], authors differentiated Arabic and Latin texts using Gabor filters. Experimental results show the capability of Gabor filters to capture script features.

B. *Features Proposed in this Work*

We propose to extract some structural features distinctive to each type of writing. In fact, structural features are intuitive aspects of writing, such as loops, branch-points, end-points and dots. They mostly affect mostly the physical structure of words. Some structural features such as PAWs, ascenders, descenders, loops and upper and lower diacritic points considering their position in the word are already used in [13] to identify printed Arabic and Latin scripts. Here, we propose to test with some new structural features which include:

- Presence of bottom diacritic points: There are no diacritic points at the bottom in Latin words which is not the case of Arabic script (see Fig. 12).

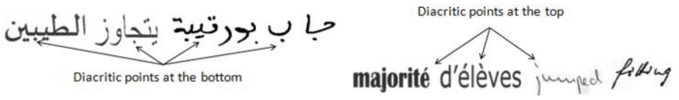

Fig. 12. Diacritic positions.

- Loop position: Loops in Arabic are generally written in the central band of the word with the exception of one Arabic letter in which the loops protrude slightly above and below. In Latin, there are a lot of letters which have loops above and below the central band (see Fig. 13).

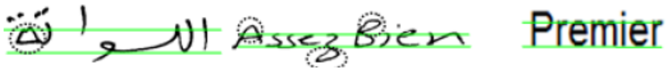

Fig. 13. Loop position.

- Presence of elongate descenders: Arabic script, whether printed or written, is characterized by the frequently presence of elongated descenders. In Latin script, descenders tend to be vertical (see Fig. 14).

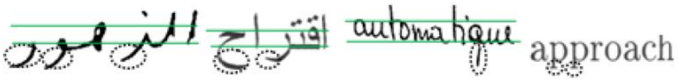

Fig. 14. Descender shape.

Because some of studied or proposed features could be unnecessary or even redundant, their suitability should be analyzed. Feature selection aims to find the best subset of features that perform better than the original ones, and also, results in a more efficient classifier. There are two main groups of methods for feature selection: feature set searching and linearly combining features for getting lower dimensionality. Among search methods, we used *BestFirst* [20], Genetic algorithm [19] and *Ranking* [18]. For feature combination, we proceeded by the *Principal Component Analysis* (*PCA*) method [17] which is one of the main classical methods for reducing dimensionality.

V. Experimental Results Analysis

Experiments have been carried using two public databases: IAM database for Latin handwritten and IFN-ENIT for Arabic handwritten words. For Latin and Arabic machine-printed scripts, we created our own database by extracting words from various magazines and newspapers which contain variable font styles and sizes. A scanning resolution of 300dpi is employed for digitization of all the words (see Fig. 15).

The training and test words have 1720 samples each, consisting of equal number of Printed Arabic (PA), Handwritten Arabic (HA), Printed Latin (PL) and handwritten Latin (HL) words. In the training phase, the features and the correct classification are used.

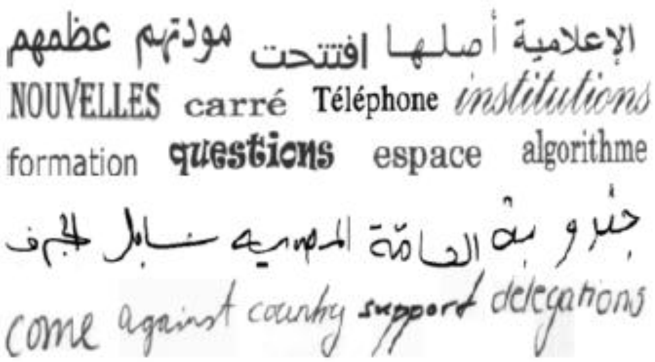

Fig. 15. Some used words.

For the test, we used the cross validation method: the words were divided into ten non-overlapping sets. Each time a classification model was calculated with training examples taken from nine sets and evaluated on the remaining sets. This procedure was repeated ten times, each time using a different set as training examples. In Table II, we give the correct identification rates for each proposed and previously used feature using *Bayes* (*AODEsr* classifier) [21].

TABLE II. Features set after selection.

Feature Set	***Precision***	***Recall***	***Measure***
Vertical projection variance	0.43	0.52	0.45
Connected component width, height, aspect ratio, area, density	0.81	0.81	0.81
Separator length between two successive connected components	0.57	0.56	0.56
Connected component profiles analysis	0.61	0.50	0.46
Loop ratio	0.53	0.53	0.52
Pixels distribution	0.32	0.42	0.36
Baseline profile features	0.58	0.84	0.85
Run length histogram	0.90	0.89	0.89
Crossing count histogram	0.73	0.73	0.72
Hu Moments	0.58	0.56	0.57
Central Moments	0.25	0.29	0.23
Upper lower profile	0.52	0.52	0.51
Bi-level Co-occurrence	0.46	0.48	0.43
Overlapping areas	0.59	0.59	0.58
Physical sizes	0.75	0.74	0.74
Steerable pyramid transform	0.66	0.66	0.66
Gabor filters	0.89	0.89	0.89

By combining these features with those proposed features in this work, correctly classified instances are 1696 using AODEsr. Only 24 are incorrectly classified. So with these 126 features, words can be identified, in a reliable way, with a correct identification rate of almost 98.60%. The time taken to build model is 0.31 seconds. Table III displays the obtained results with different feature selection methods using *AODEsr* classifier. It also indicates the time taken to build model.

TABLE III. Testing with different feature selection methods.

Evaluator	***Search***	***Selected features***	***Identification Rate (%)***	***Time (s)***
CfsSubsetEval	GeneticSearch	58	98.72	0.03
CfsSubsetEval	BestFirst	36	98.43	0.03
PCA	Ranker	48	95.69	0.02

As shown in Table III, when applying the GeneticSearch as feature set selection method, the selected features are reduced from 126 to 58 features (see Table IV), the correctly identification rate is the highest and the consuming time is among the lowest. Notice that *CfsSubSetEval* [21] evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.

TABLE IV. FEATURES SET AFTER SELECTION.

Feature Set	***Selected Features***
Vertical projection variance	(1/1)
Connected component width, height, aspect ratio, area, density	(2/10)
Separator length between two successive connected components	(2/2)
Connected component profiles analysis	(1/1)
Loop ratio	(1/2)
Pixels distribution	(1/1)
Baseline profile features	(5/8)
Run length histogram	(9/15)
Crossing count histogram	(2/10)
Hu Moments	(2/2)
Central Moments	(1/2)
Upper lower profile	(2/3)
Bi-level Coccurrence	(2/4)
Overlapping areas	(1/1)
Physical sizes	(2/5)
Steerable pyramid transform	(19/48)
Gabor filters	(2/8)
Proposed features	(3/3)

As results on Table V show, the average accuracy is the same, about 98.72% for handwritten and machine printed words either in Arabic or Latin scripts. Notice that, with the selected features, printed Arabic words can be identified, in a reliable way, with a correct identification rate of 100%.

TABLE V. DETAILED ACCURACY BY CLASS.

	PA	***HA***	***PL***	***HL***	***Average***
Precision	0.982	0.991	0.998	0.979	0.987
Recall	1	0.977	0.995	0.977	0.987
F-Measure	0.991	0.984	0.997	0.978	0.987

When observing the confusion matrix (see Table VI), we note that it is about confusion cases between handwritten Arabic and Latin scripts. Most of them mainly come from their cursive nature. Confusion, between printed and handwritten Latin script arise because of the writing styles of many writers who do not use ligatures between the letters.

TABLE VI. CONFUSION MATRIX.

	PA	***HA***	***PL***	***HL***
PA	430	0	0	0
HA	3	420	0	7
PL	0	0	428	2
HL	5	4	1	420

We also compared the performance of three typical classifiers: *Bayes* (*AODEsr*) [23], *k-Nearest Neighbor* (*k-NN*)[22] and *Decision Tree* (*J48*)[23] (see Table VII).

TABLE VII. ACCURACY BY CLASSIFIER.

Classifier	***F-Measure***
AODEsr	98.72%
J48	85.98 %
k-NN	97.5%

TABLE VIII.

In Fig. 16, we display the *Receiver Operating Characteristics* (*ROC*) curve to compare the three classifiers and to highlight what is the classifier that has the best discriminative power. This will be the classifier that has the highest *ROC* curve widening. Here, it corresponds to *AODEsr* considering HL class.

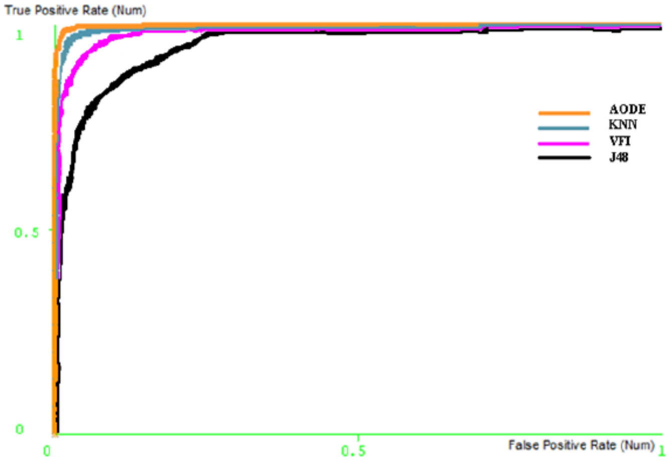

Fig. 16. Comparing with Existing System

Using the selected features, the *AODEsr* classifier captures significant amount of the differences between machine-printed and handwritten Arabic and Latin words providing a good solution for this task. To compare it with a system proposed in [12] which deals with the same problem, we used a common database (our database composed of 1720 word samples). Recall that [12] proposed an identification system, at word level, based on steerable pyramid transform and using *k-NN* as classifier. Table VIII summarizes the obtained results.

System	*Feature set*	*Classifier*	*Data-Base*	*F-Measure*
[12]	48	k-NN	800	97.5%
[12]	48	k-NN	1720 (common database)	93.08%
[12]	48	AODEsr	1720(common database)	91.86%
Our system	58	AODEsr	1720(common database)	98.72%

Notice that when testing [12] system on a larger database (1720 words instead of 800), the *F-Measure* is reduced from 97.5% to 93.08%. But the use of *AODEsr* instead of *k-NN* as classifier has further reduced the rate to 91.86%. In sum, the use of a set of 58 selected features with *Bayes* classifier achieves an identification rate of 98.72% which is slightly better than 93.08% (the identification rate obtained using features from the steerable pyramid transform with k-NN as classifier and tested on the same database). In our view, the obtained results show that giving slightly higher weight to the structural information can produce better results.

VI. Conclusion

In this work, we aimed to introduce and survey script and nature identification at word level. We carried experiments on Arabic and Latin handwritten and machine-printed words and tried to propose and select features that maximize the distinction between words. We retained a set of 58 simple and different features after selection. We compared the performances of three classifiers. We used standard databases for testing the proposed features and classifiers and the results show the identification process is robust and reliable at the word level. Notice that these features may be generalized to include all other Romance and Anglo Saxon languages instead of only English or French and other languages that use Arabic scripts such as Persian and Urdo. In the future, we plan to explore further features and classifiers.

References

[1] A. Kacem, A. Saïdani and A. Belaïd, “A System for an automatic reading of student information sheets”, Proc. of ICDAR, 1265-1269, (2011).

[2] A. Saïdani, A. Kacem and A. Belaïd, “Identification of machine-printed and handwritten words in Arabic and Latin scripts”, Proc. of ICDAR, (2013)

[3] S. Kanoun, I. Moalla, A. Ennaji and A. M. Alimi , “Script Identification for Arabic and Latin Printed and handwritten Documents”, Proc. of DAS, 159-165, (2000).

[4] S. Mozaffari and P. Bahar, “Farsi/Arabic handwritten from machine-printed words discrimination”, Proc. of ICFHR, 694-699,(2012).

[5] L. Faria da Silva, A. Conici and A. Sanchez, “Automatic discrimination between printed and handwritten text in documents”, Proc. of SIBGRAPI, XXII Brazilian Symposium, 261-267, (2009).

[6] L. Zhou , Y. Lu and C. Tan , “Bangla / english Script identification based on Analysis of Connected Components Profiles”, Proc. of DAS,243-254,(2006).

[7] M. Ben Jlaiel , S. Kanoun, A. M. Alimi and R. Mullot , ”Three decision levels strateg for Arabic and Latin texts differentiation in printed and handwritten natures”, Proc. of ICDAR, 1103-1107, (2007).

[8] E. Kavallieratou and S. Stamatatos, “Discrimination of machine-printed from handwritten Text using Simple Structural Characteristics”, Proc. of ICPR, 437-440, (2004).

[9] Y. Zheng , C. Liu and X. Ding , ”Single character type identification”, Proc. of DRR, 49-56,(2002).

[10] A. M. Elgammal and M. A. Ismail, “”Techniques for Language Identification for Hybrid Arabic-English Document Images”, Proc. of DRR, 1100-1104, (2001).

[11] K. Baâti , Kanoun and M. Benjlaiel , ”Diffrenciation d’criture arabe et Latine de natures imprime et Manuscrite par approche globale”, Proc. of CIFED,(2010).

[12] M. Benjelil , R. Mullot and M. A. Alimi, ”Language and script identification based on Steerable Pyramid Features”, Proc. of ICFHR, 18-20 September, Bary-Italy, 712-717,(2012).

[13] S. Haboubi , S. S. Maddouri and H. Amiri , ”Discrimination between Arabic and Latin from bilingual documents”, Proc. of CCCA,(2011).

[14] G. G. Rajput and H. B. Anita, “handwritten Script Identification from a Bi-Script Document at Line Level using Gabor Filters”, Proc. of SCAKD, (2011).

[15] Y. Zheng , H. Li and D. Doermann , ”Machine-printed text and handwritten identification in noisy documents images”, IEEE PAMI, 26(23):337–353, (2004).

[16] H. Ma and D. Doermann , ”Word level script identification for scanned document images”, Proc. of ICDAR,(2004).

[17] R. O. Duda , P. E. Hart and D. G. Stroke , ”Pattern classification”, Wisley interscience,(2001).

[18] B. W. Cavnar and J. M. Trenkle, ”N-gram based text categorization”, Proc. of SDAIR, 161-175,(1994).

[19] D. E. Golberg , ”Genetic algorithm in search, optimization and machine learning”, Addison-wesley, (1989).

[20] http://wiki.pentaho.com/display/DATAMINING/BestFirst

[21] M. A. Hall , ”Correlation-based Feature Subset Slection for Machine Learning”, Hamilton, New Zealand,(1998).

[22] T. M. Cover and P. E. Hart , ”Nearest neighbor pattern classification”, IEEE Trans. Information Theory, 13(1), 21-27,(1967).

[23] http://www.d.umn.edu/ padhy005/Chapter5.html.

[24] S. Haboubi , S. Snoussi Maddouri and N. Ellouze , ”Diffrenciation de documents textes Arabe et Latin par filtre de Gabor”, Proc. of TAIMA, (2007).