Chapter 7

A Novel Method for Multifont Arabic Characters Features Extraction

Nadia Ben Amor and Najoua Essoukri Ben Amara

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/52245

1. Introduction

Recently, many researchers around the world focused on Arabic document analysis, promising results have been reported.

However, there are not standard databases in Arabic to be considered as a benchmark. Each of research groups implemented their own system of set of data they gathered and different recognition rates were reported. Therefore, it is very difficult to give comparative and objective results for the proposed methods.

The aim of our work is to test several feature extraction algorithm and classification method using the same data base that we developed and which is composed of some 664 488 Arabic characters in nine different fonts and to conclude as far as the best suitable method for Arabic morphological specificities.

2. A review of Arabic characteristics

In this section we present a description of the important aspects of Arabic characters since the characteristics of Arabic writing is different from other alphabets.

Arabic script is cursive in both its handwritten and printed forms and letter shape is context sensitive.

The cursive nature of Arabic script is the main challenge to any Arabic text recognition system. Besides, Arabic script cursiveness obeys well-defined rules: some letters of the alphabet are never connected to their successors while others link to their within-word successors by a horizontal connection line.

In addition to the cursive aspect, we can also note the multitude of directions that can be described by the same Arabic character, especially in the multifont context.

Arabic writing may be classified into three different styles [1, 9]:

- Typewritten: This is a computer generated style. It is the simplest one because the characters are written without ligature or overlaps Figure1.

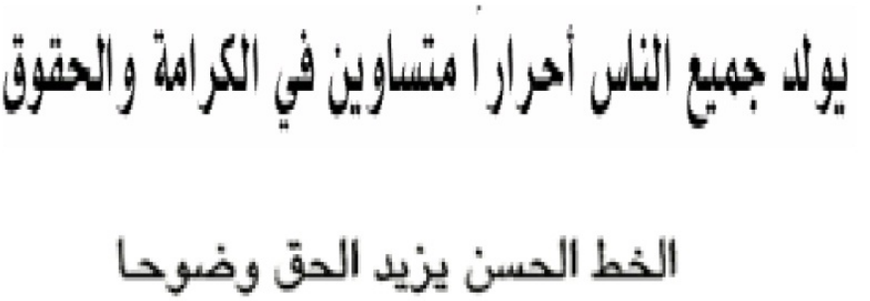


Figure 1. Example of typewritten Arabic style

- Typeset: This style is more difficult than typewritten because it has many ligatures and overlaps. It is used to write books and newspapers.

Nowadays, this style may also be generated using computers Figure2.

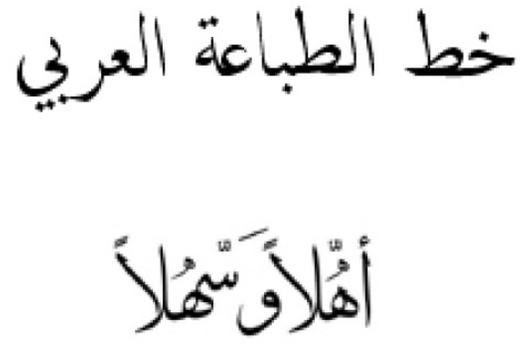


Figure 2. Example of typeset Arabic style

- Handwritten: This is the most difficult style because of the variation of writing the Arabic alphabets from one writer to another.

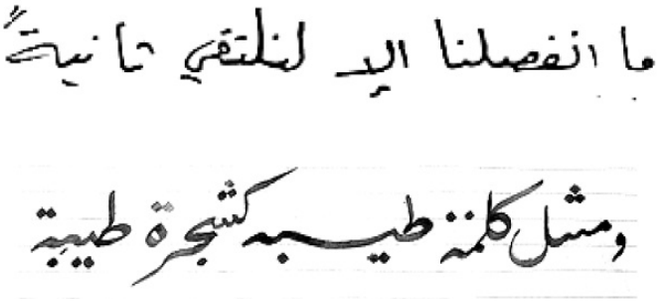


Figure 3. Examples of handwritten Arabic

Besides to different style of writing, there are many fonts in Arabic which make the recognition process more and more difficult.

In our work we have been dealing with multifont Arabic isolated characters. In fact, Segmenting Arabic script into characters is very difficult and always generates errors in the segmentation–based system. This work solves the cursiveness problem by presenting a segmentation–free system.

Due to the lack of common Arabic script data base, we had to develop our own one including all the shapes of the Arabic characters, beforehand segmented.

These characters was considered in nine different fonts which are Arabic transparent, Badr, AlHada, Diwani, Kufi, Cordoba, Andalus, Ferisi and Salam (Figure 4, Figure 5).

Characters / Fonts	Mim	Té	Noun	Lèm	Sin
Arabic Transparent	م	ت	ن	ل	س
Badr	م	ت	ن	ل	س
Diwani	م	ت	ن	ل	س
Kufi	م	ت	ن	ل	س
AlHada	م	ت	ن	ل	س
Andalus	م	ت	ن	ل	س
Cortoba	م	ت	ن	ل	س
Ferisi	م	ت	ن	ل	س
Salam	م	ت	ن	ل	س

Figure 4. Samples of isolated Arabic characters considered in nine different fonts.

Besides, theses characters were considered in the different shapes they could have depending on their position within a word. Some samples of these different shapes are represented in the figure 5.

In fact, more and more Arabic documents are compound and use the multifont context, such as the newspapers and the magazines or even the official documents. Figure 6, extracted from an official Tunisian Newspaper, includes three different fonts which are Arabic Transparent, Ferisi and Andalus, used in the big title and the subtitles.

Characters / Fonts	Noun	He	Ta	Sad	Fe	Ain	Kaf	Ke
Arabic Transparent	ن نن	ه هه	ط طط	ص صص	ف ففف	ع عع	ق ققق	ك ككك
Badr	ن نن	ه هه	ط طط	ص صصص	ف ففف	ع عع	ق ققق	ك ككك
Diwani	ن نـنـن	ه هـهـه	ط ططط	ص صصص	ف فـفـف	ع عـع	ق قـقـق	ك كـكـك
Kufi	ن نن	ه هه	ط طط	ص صصص	ف ففف	ع عع	ق ققق	ك ككك
AlHada	ن نن	ه هه	ط طط	ص صصص	ف ففف	ع عع	ق ققق	ك ككك
Andalus	ن نن	ه هه	ط طط	ص صصص	ف ففف	ع عع	ق ققق	ك ككك
Cortoba	ن نن	ه هه	ط طط	ص صصص	ف ففف	ع عع	ق ققق	ك ككك
Ferisi	ن نـنـن	ه هـهـه	ط ططط	ص صصص	ف فـفـف	ع عـع	ق قـقـق	ك كـكـك
Salam	ن نن	ه هه	ط طط	ص صصص	ف ففف	ع عع	ق ققق	ك ككك

Figure 5. Samples of different Arabic characters shape according to their font and position in a word.

Figure 6. Examples of Arabic multifont documents, extracted from two official newspapers

We have developed so far several processes for mutltifont Arabic characters recognition. All of these methods have proved the importance of the cooperation of different types of information at different levels (feature extraction, classification, post-processing…). This cooperation helps to overcome the variability of Arabic script especially in a multifont context [12, 13, 14, 15].

In this paper we highlight the role of Contourlets in the feature extraction step in an Arabic OCR context. This will allow us to compare the Contourlets performances with those of Wavelets and the Standard Hough Transform (SHT) that we previously used for the same purpose in our multifont Arabic recognition system. This comparison will lead to conclude as far as the precious contribution of the Contourlets in Arabic characters recognition field.

In the following section, we present the first approaches we developed in the features extraction step, then we introduce the Contourlet transform in the 3rd section. In section 4, we detail the system performances and experimental results. And finally, we conclude this paper in Section 5.

3. Wavelets and SHT approach for Arabic characters feature extraction

Feature extraction is a preliminary step for characters recognition. However, there is no perfect edge detector or feature extraction algorithm.

Many approaches have been so far developed for many alphabets such as Latin and Japanese. Yet given the specificity of this kind of writing we cannot apply them, as they are, for Arabic characters. Indeed, Arabic writing presents a very specific morphology. Thus the field remains one of the most challenging even though some works have been done [6, 17].

Arabic script is mainly composed of graphemes of cursive and structural nature. That's why we developed first two approaches based on wavelets transform and standard Hough transform-SHT. Wavelet transform is suitable for extracting cursive characteristics, while SHT is well known for extracting directional features.

Even though these methods have allowed us to achieve good recognition rates, it is worth mentioning that they presented some weaknesses regarding the pure directional and cursive aspect of some Arabic characters such as....

In fact, the wavelet transform has been proven to be powerful in many signal and image processing applications such as compression [11], noise removal, image edge enhancement and feature extraction.

However, wavelets are not optimal in capturing the two-dimensional singularities found in images. They are not effective in representing the images with smooth contours in different directions even though they offer multi-scale and time-frequency localization of an image (Figure7, Figure8). Wavelets are known to be quite efficient in representing image textures, but they show up insufficient as far as the smooth contour localization is concerned [16].

Typically, a separable 2-D wavelet transform provides:

- multiresolution, which is the ability to visualize the transform with varying resolution from coarse to fine

- localization, which is the ability of the basis elements to be localized in both the spacial and frequency domains
- critical sampling, which is the ability for the basis elements to have little redundancy.

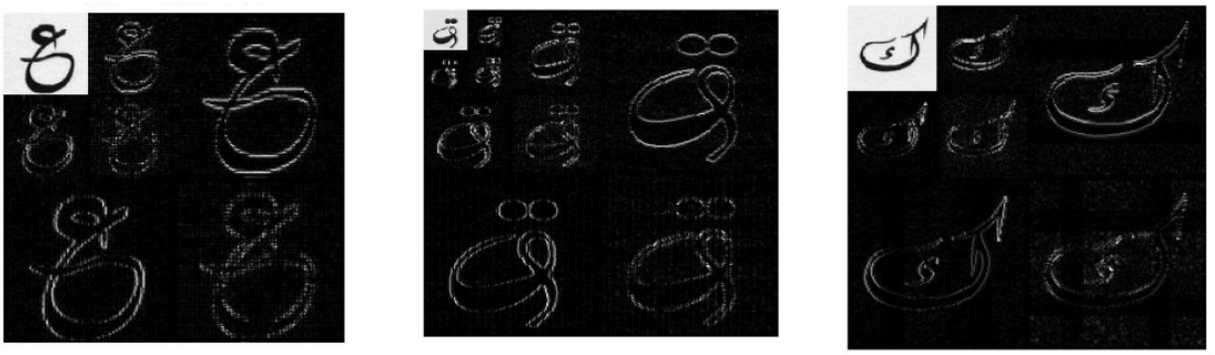


Figure 7. Examples with good recognition results using wavelets as feature extractor (cursive aspect)

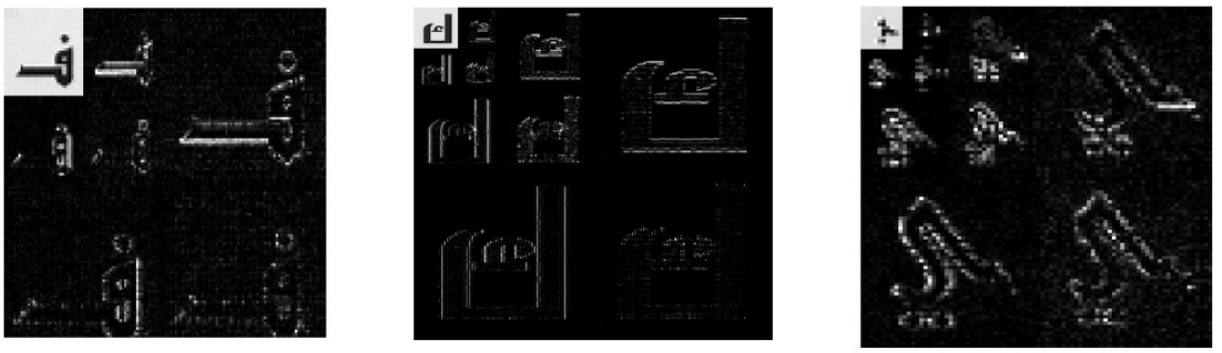


Figure 8. Examples with less good recognition results using wavelets as feature extractor (directional aspect)

However, it is not capable of providing:

- directionality, which is having basis elements defined in a variety of directions
- anisotropy, which is having basis elements defined in various aspect ratios and shapes.

In fact, despite its efficiency the wavelet transform can only capture limited directional information. This can affect the performance of the recognition system especially that the cursive nature of Arabic characters leads to a large number of directions to be considered. Thus the introduction of a directional based feature extraction method was a necessity.

The other features extraction method we focused on, was the SHT.

The SHT is known to be the popular and powerful technique for finding multiple lines in a binary image, and has been used in various applications.

It is very useful when dealing with the identification of features of a particular shape within a character image such as straight lines, but it fails as soon as it's a question of curves and circles localization [9]. This fact is shown in Figure9 and Figure10.

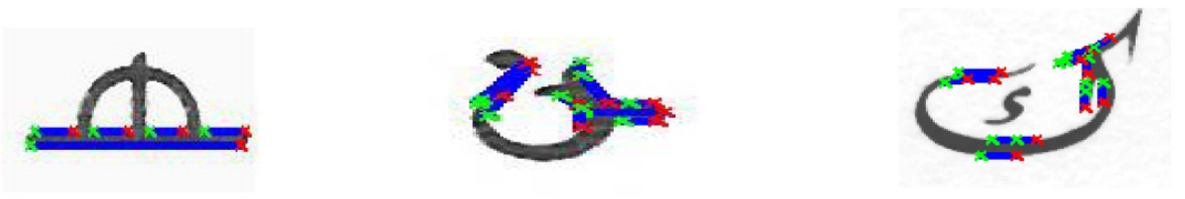


Figure 9. Examples of characters where the SHT fails in capturing cursive forms

Figure 10. Examples of characters where the SHT manages in capturing straight forms

Besides, trying to take advantage of these two previous methods, we have integrated them in a hybrid approach. This hybridization allowed localizing image texture as well as straight lines and directional features. In spite of the improvement of the results, the computation time had considerably increased [14].

4. Discrete Contourlet transform and feature extraction

Recently, several transforms have been proposed for image analysis that have incorporated directionality and multi-resolution which could more efficiently capture edges in the processed images. In fact, much more elaborated techniques of signal processing emerged such as Steerable Pyramid [4], Curvelets [3] and Contourlets [7] which are some well known examples. The Contourlet transform is one of the new geometrical image transforms, which seems to be promising since it allows extracting both directional and cursive primitives.

The contourlet transform uses a stage of subband decomposition followed by a directional transform. In the contourlet transform, a Laplacian pyramid is applied in the first stage, while directional filter banks (DFB) are used in the angular decomposition stage [7].

Unlike Wavelets, the contourlet transform is a directional transform capable of capturing contours and fine details in images.

In addition, the contourlet expansion is composed of basis functions oriented at a variety of directions in multiple scales. With this rich set of basis functions, the contourlets can effectively capture smooth contours.

Contourlets not only possess the main features of wavelets (multiscale and time-frequency localization), but also offer a high degree of directionality and anisotropy. Precisely, Contourlets transform involves basis functions that are oriented at any power of two's number of directions with flexible aspect ratios. [8]

The double filter bank structure of the contourlet is shown in Figure 11 for obtaining sparse expansions for typical images having smooth contours.

4.1. Laplacian Pyramid decomposition

The first filter bank, known as the Laplacian Pyramid (LP), is utilized to generate a multiscale representation of an image of interest. LP decomposition at each level generates a down-sampled low-pass version of the original image and the difference between the original and the prediction, which results in a band-pass image. The LP decomposition is shown in Figure 12. In LP decomposition process, H and G are one dimensional low pass analysis and synthesis filters respectively. M is the sampling matrix. Here, the band-pass image obtained in LP decomposition is then processed by the directional filter bank stage to reveal the directional details at each specific scale level.

The output values from the second filter bank are called "contourlet coefficients". Any analysis performed with the contourlet coefficients is considered as in the "contourlet domain."

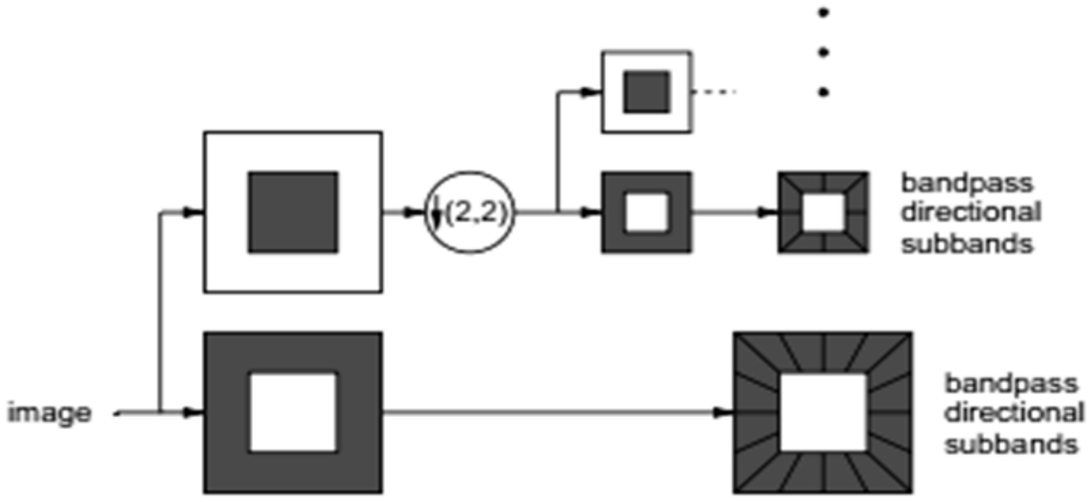


Figure 11. Double Filter Bank Decomposition of Contourlets transform.

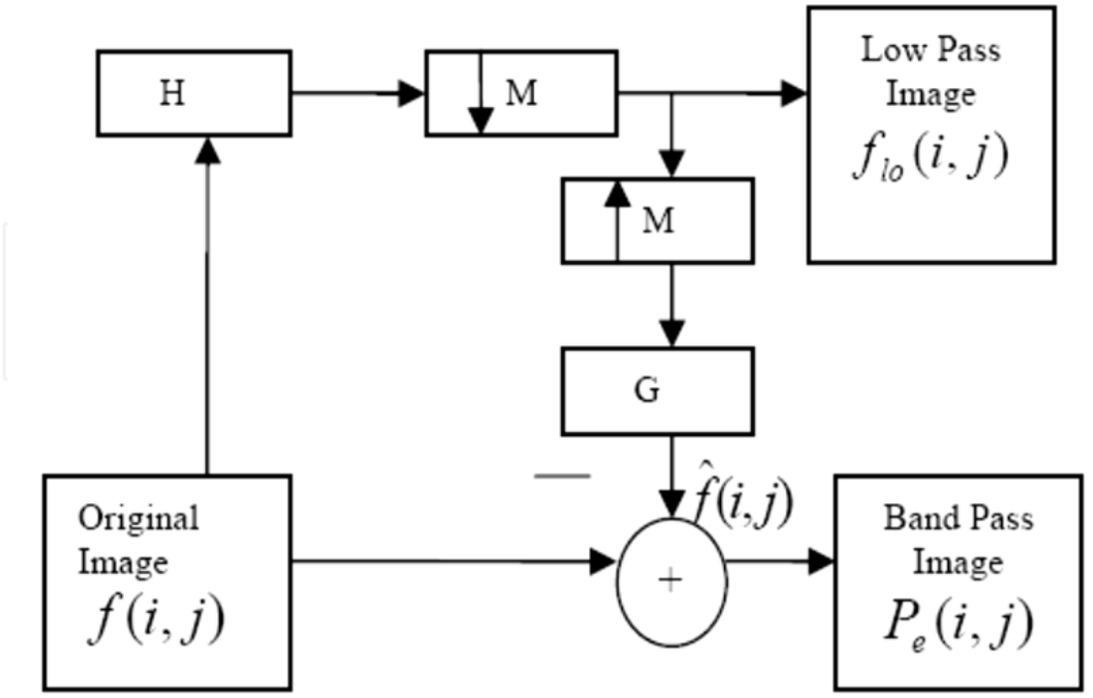


Figure 12. The principle of LP

4.2. Directional Filter Bank decomposition

Directional Filter Bank (DFB) is designed to capture the high frequency content like smooth contours and directional edges. Several implementations of these DFBs are available in the literature [8]. Combination of a Laplacian Pyramid (LP) and a DFB gives a double filter bank structure known as contourlet filter bank. Band pass images from the LP are fed to DFB so that directional information can be captured.

The scheme can be iterated on the coarse image. This combination of LP and DFB stages results in a double iterated filter bank structure known as contourlet filter bank, which decomposes the given image into directional sub-bands at multiple scales.

Since the purpose of using Contourlets is to focus on the cursive nature of the Arabic characters, we take an example of a cursive area and examine the behaviour of both wavelets and Contourlets on it Figure13.

Figure.13.a shows how wavelets arrange each others along the edge at different resolutions. The small blue squares represent the wavelets at the finest resolution, the green ones represent intermediate resolution and the red squares represent wavelets at the coarsest resolution. Figure.13.b shows the alignment of Contourlets and we can notice that the squares are replaced by rectangles.

Besides, we notice that, at each resolution, the edge can be represented by a far less number of contourlets than wavelets. As Wavelets are isotropic they can not take advantage of the underlying geometry of the edge. They approximate the edge as a collection of dots (small

squares) so many points are needed to represent an edge. While contourlets are representing the edge as a collection of small needles hence only a few needle shaped line segments can represent the edge.

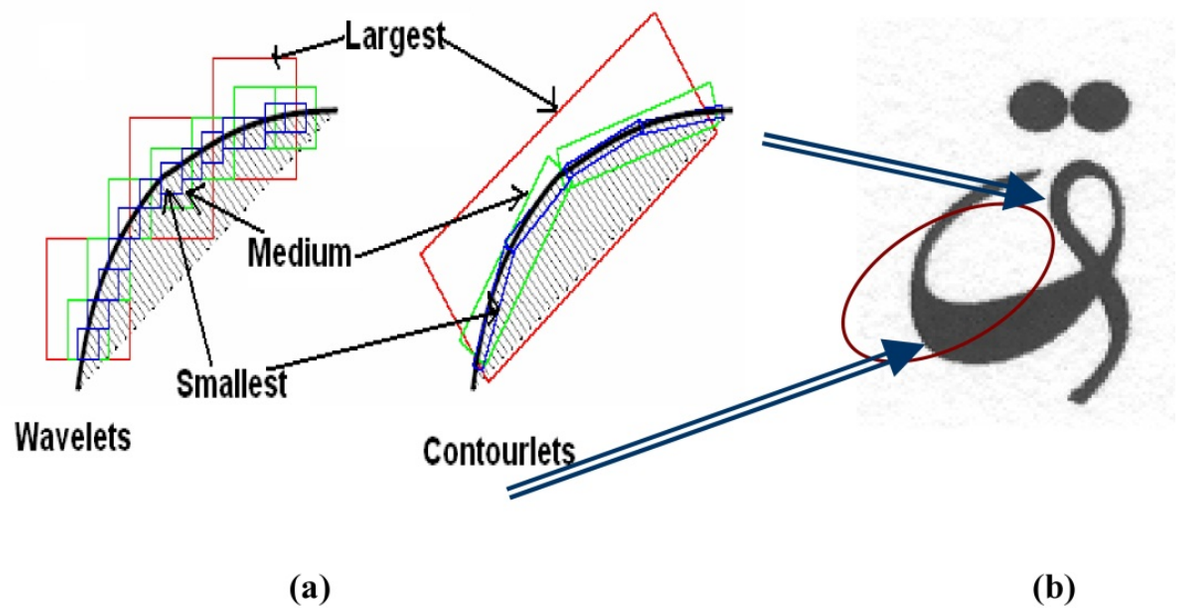


Figure 13. Wavelets (a) vs. Contourlets (b)

To sum up, one contourlet may be assumed to be formed by grouping several wavelets at the same resolution.

In the Figure 14, we present some examples of Arabic characters images decomposition, using Contourlets, wavelet and SHT. The better quality comparing with Wavelets and SHT is obvious.

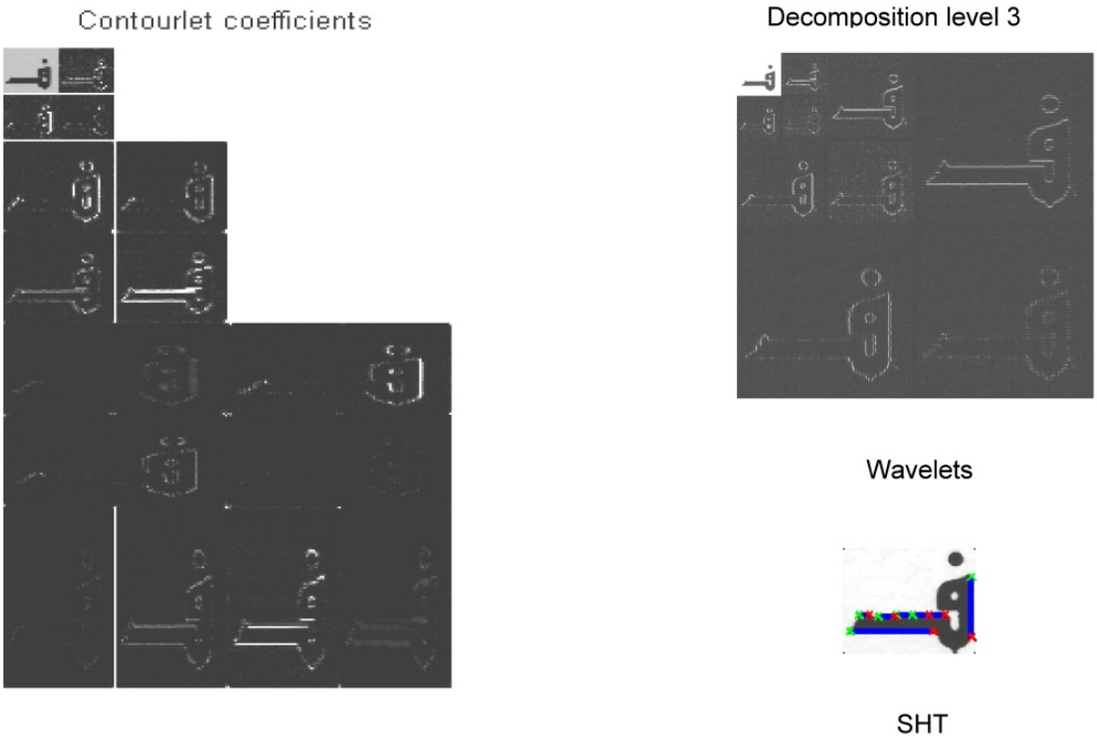
5. Experimental results

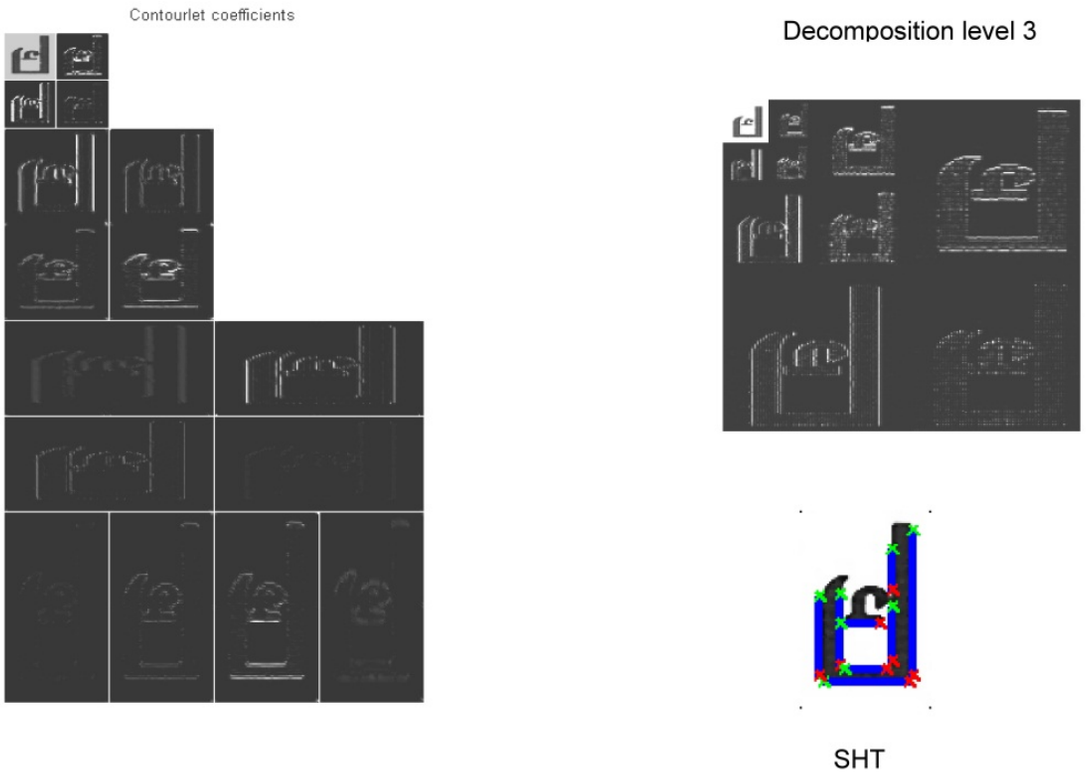
Due to the lack of a standard database in Arabic to be considered as a benchmark, we developed our own database including all the Arabic characters beforehand segmented and presented in the different shapes they could have in a word.

All images in the database are processed in the grey level in the Tiff format.

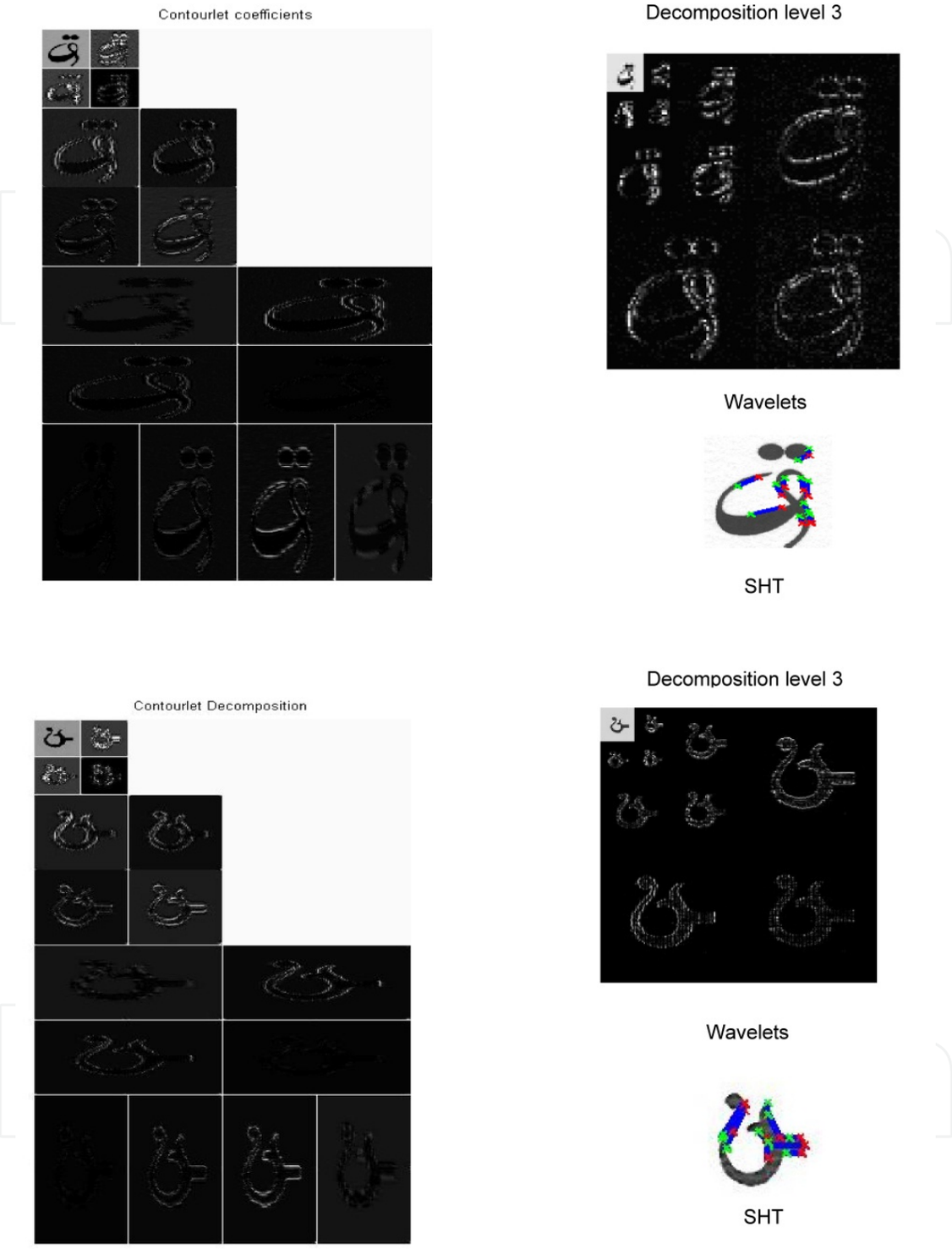
Each image is decomposed in the contourlet domain. The resulting coefficients are structured in a special cellular form. Many experiments were conducted and we retained the Standard Deviation (SD) vector as a set of features.

Edge and texture orientations are captured by using contourlet decomposition with 3 level (0, 2 and 3) decomposition. At each level, the numbers of directional subbands are 3, 4 and 8 respectively. 'Pkva' filters are used for LP decomposition and directional subband decomposition.





(a) Better straight lines and directions detection than wavelets and SHT.



(b) Better curves and directions detection than wavelets and SHT.

Figure 14. Examples of images of features extraction using Contourlets, Wavelets and SHT: Better quality and recognition rates than Wavelets at greater level of resolution. Better curves detection than SHT.

As a result of this process, we obtain as output, a cell-vector where except output {1} corresponding to the lowpass subband. Each cell corresponds to one pyramidal level and is a cell-vector that contains band-pass directional subbands from the DFB at that level. These parameters result in a 16-dimentional feature vector (n=16). Standard deviation vector used as image feature is computed on each directional sub-band of the contourlet decomposed image and then normalized. This normalized feature vectors are used to feed the entry of the Artificial Neural Network classification stage.

Two architectures of neural network were implemented: a global Multilayer Perceptron (MLP) and a modular one.

In Table 1, we present the different recognition rates achieved when using Contourlets [13], Wavelets [10] and SHT [11] in features extraction. These results show the efficiency of contourlet transform compared to those obtained previously with the SHT and wavelet transform even though the used directional filter is a predefined one.

Features extraction / Classification		Contourlets		Wavelets		SHT
Characters		Modular MLP	Global MLP	Modular MLP	Global MLP	Modular MLP
ا	Alif	99.43	99.58	99.52	97.70	99.10
ب	Ba′	99.41	99.85	99.43	98.85	95.13
ت	Ta′	98.66	99.56	98.56	96.79	97.16
ث	Tha′	99.18	99.31	99.04	98.75	98.86
ج	Jim	99.28	100	97.99	97.50	97.02
ح	Ha′	99.90	100	99.43	98	96.80
خ	Kha′	99.40	100	98.85	97.23	96.26
د	Dal	99.27	99.73	100	98.37	95.69
ذ	Dhal	98.43	99.60	98.18	97.96	96.73
ر	Ra′	98.84	99.51	97.89	96.79	96.73
ز	Zay	98.57	99.74	95.41	95.08	96.55
س	Sin	99.15	99.85	99.52	98.18	94.65
ش	Chin	98.59	99.90	98.76	97.75	96.16
ص	Sad	98.35	99.83	99.52	100	96.24
ض	Dhad	97.96	99.71	96.47	96.55	94.62

Features extraction / Classification		Contourlets		Wavelets		SHT
Characters		Modular MLP	Global MLP	Modular MLP	Global MLP	Modular MLP
ط	Ta′	98.83	99.76	98.37	98.23	94.64
ظ	Dha′	98.87	99.22	98.66	97.85	95.24
ع	‘Ayn	99.23	99.82	98.47	95.61	96.76
غ	Ghayn	99.18	99.78	98.76	98.63	96.22
ف	Fa′	98.29	99.20	99.23	100	95.50
ق	Qaf	99.69	99.64	98	98.97	95.34
ك	Kaf	99.02	99.71	99.52	98.69	95.16
ل	Lam	99.45	99.60	100	98.93	98.38
م	Mim	99.24	99.47	99.33	99.16	97.02
ن	Noun	99.19	99.42	97.51	96.92	97.18
ه	Ha′	98.40	99.85	98.76	98.37	98.55
و	Waw	99.31	98.98	98.95	97.86	97.31
ي	Ya′	99.66	99.88	98.09	98.12	97.98
Average rate (%)		99.03	99.66	98.65	97.95	96.53

Table 1. Recognition rate per character corresponding to the MLP models

6. Conclusion and perspectives

In this paper, we were interested in the challenges of Arabic characters feature extraction especially in a multifont context. We proposed a new approach for Arabic character recognition based on contourlet transform for feature extraction.

The achieved results show the efficiency of this transform compared with the Wavelet transform and the SHT. They proved its superiority in describing the different morphological variations of Arabic isolated characters. In fact, the contourlet transform have the advantage of highlighting both directional and cursive nature of Arabic scripts.

As a major perspective to this work we can consider to optimize the Contourlets Algorithm, by developing an adaptive filter depending on the character’s class and form. Such as implementing filters adapting the most recognized directions by the SHT and of course the main directions of the Arabic scrip itself.

Author details

Nadia Ben Amor
National Engineering School of Tunis, Country

Najoua Essoukri Ben Amara
National Engineering School of Sousse, Country

7. References

[1] B. Al-Badr and S. A. Mahmoud. Survey and bibliography of Arabic optical text recognition. Signal Processing, 41(1):49–77, 1995.

[2] D.Y .Po and Minh N Do. "Directional Multiscale Modelling of Images Using the Contourlet-transform", IEEE Transactions on Image Processing, 2006 Vol. 15, No. 6, pp 1610- 1620.

[3] E. J. Candes and D. L. Donoho, "Curvelets – a suprizingly effective nonadaptive representation for objects with edges," in Curve and Surface Fitting, Saint- Malo, Vanderbuilt Univ. Press, 1999.

[4] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation" 2nd IEEE International Conference on Image Processing.Washington, October, 1995DC. vol III, pp 444-447.

[5] E. W. Brown, "Character Recognition by Feature Point Extraction", Northeastern University internal paper 1992.

[6] M.Hamdani , H. El Abed, M. Kherallah, and A. M. Alimi, " Combining multiple HMMs using online and offline features for offline Arabic handwriting recognition," In Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 201–205, July 2009.

[7] M. N. Do and M. Vetterli, "Contourlets", in Beyond Wavelets, Academic Press, New York, 2003.

[8] M. N. Do, "Directional multiresolution image representation", Ph.D. Thesis. Department of Communication Systems, Swiss Federal Institute of Technology Lausanne, November 2001.

[9] M. S. Khorsheed. Off-line arabic character recognition - a review. Pattern Analysis & Applications, 5:31–45, 2002.

[10] N Aggarwal and WC Karl. Line detection in images through regularized Hough transform. IEEE Trans. on Image processing, 15:582–591, 2006.

[11] N.Ben Amor, N. Essoukri Ben Amara "DICOM Image Compression By Wavelet Transform". Proc. IEEE International Conference on Systems, Man and Cybernetics, Vol. 2, 6-9 October 2002 Hammamet, Tunisie.

[12] N.Ben Amor, N. Essoukri Ben Amara "Applying Neural Networks and Wavelet Transform to Multifont Arabic Character Recognition" International Conference on

Computing, Communications and Control Technologies (CCCT 2004), Austin (Texas), USA, on August 14-17, 2004.

[13] N.Ben Amor, N. Essoukri Ben Amara "Multifont Arabic Characters Recognition Using Hough Transform and Neural Networks" the Third International Symposium on Neural Networks (ISNN 2006) Chengdu China,May 28-31, 2006, J. Wang et al. (Eds.): (ISNN 2006), Lecture Notes in Computer Sciences LNCS 3972, 2006.© Springer-Verlag Berlin Heidelberg 2006, , pp. 293 – 298.

[14] [N.Ben Amor, N. Essoukri Ben Amara "A hybrid Approach for Features Selection in multifont Arabic isolated characters Recognition" the International Conference on Computer & Communication (ICCCE06), Malaysia, 9-11 May 2006.

[15] N.Ben Amor, N. Essoukri Ben Amara " Multifont Arabic Isolated Character Recognition Using Contourlets and Artificial Neural Networks" 11th International Conference on Frontiers in Handwriting Recognition 19-21 August 2008 (ICFHR 08]

[16] S.Esakkirajan, T. Veerakumar, V.Senthil Murugan, R. Sudhakar, "Image compression using contourlet transform and multistage vector quantization", *GVIP Journal*, volume 6, Issue 1, pp.19-28, July 2006.

[17] T.J. Klassen, Towards NN Recognition of Handwritten Arabic Letters, Project Report, 2001.