SCALE-SPACE APPROACH FOR CHARACTER SEGMENTATION IN SCANNED IMAGES OF ARABIC DOCUMENTS

[1] NOUREDDINE EL MAKHFI AND [2] OMAR EL BANNAY

[1]Department of electrical and computer engineering,
Laboratory of Data Transmission and Processing,
Sidi Mohamad Ben Abdallah University, EST - LTTI,
Fez, Morocco

[2]Department of computer engineering
Laboratory of Informatics, System, Electrical, Networks and Telecommunications
Hassan 1 University, ENSA-LISERT, Khouribga, Morocco

E-mail: [1]n.elmakhfi@gmail.com , [2]omarelbannay@gmail.com

ABSTRACT

The characters segmentation is an important stage for the optical character recognition in documents. In this article, we present a new method for segmenting the Arabic documents into text characters. Our method based on the scale space to retrieve the blobs forming each character in the word image. These blobs detected in appropriate scales to recover the characters and cut the junctions between the text characters. The experimental results reveal that the proposed method is encouraging despite some subdivision of characters, which mainly produced by the reconciliations exaggerated between the characters in words. This subdivision can be corrected by adding new steps to merge the character fractions in the recognition phase.

Keywords: *Digital Image/Text; Scale Space; Cursive Writing; Character Segmentation; Arabic OCR.*

1. INTRODUCTION

We are currently witnessing an increase in the text available in digital format, and the presence of a large number of Arabic printed documents as images whose content is easily searchable. However, software optical character recognition "OCR" still today find it difficult to detect the Arabic printed text, especially when it comes to low-resolution images.

Printed Arabic text shows gaps between words that are much larger than the pseudo-words of the machine-printed text. The letters in the majority of cases connected together by bonds.

Segmentation is the first step in the selection of regions of interest containing useful information for text recognition. Segmenting an image comprising Arabic text is an easy process in the case of the segmentation of lines or words. However, in the case of the character segmentation, the bonds present a handicap to the segmentation of the letters.

In this paper, we present a scale space method for character segmentation. Our method is based, firstly, on the analysis of elliptical blobs "extents blobs" in the multi-scale representation of the image space, and secondly on the separation of bonds characters. The multi- scale approach used mainly in the field of segmentation regions in images. This approach based on the theory of space scale developed by T. Linderberg [1] in computer vision. Concerning the characters segmentation, we consider the related regions representing blobs of characters in the Arabic document images at different scales. These blobs detected in suitable scales to break the bonds and to locate the characters.

Given the complexity of the Arabic script and the difficulty of character segmentation, our method allows extracting the characters and Arabic letters even at low resolutions to solve problems related to the recognition of printed Arabic text.

The main advantages of using our method include reduction of noise effects and character segmentation low resolution.

2. RELATED WORK

Most studies on the segmentation of cursive writing characters based on the extraction of feature forms. We can cite the main methods of segmentation in the literature:

The methods used for the segmentation of characters in the manuscripts based on skeleton [2], [3], contour [2], [3] and sliding windows [3], [6]:

- Segmentation from the skeleton: based on calculations of the angles of curves to define the characteristic areas to extract the points of cuts. X.Dupré [3] demonstrates that the method has some delicate adjustments, and it does not give good results in the case of overlapping letters.
- Segmentation from the contour S. Madhvanath et al [5] determine candidate break points grapheme based on local extrema of the contour that are associate according to a proximity criterion. This method also requires configurations. X. Dupré [4] states that the configurations are easy if the quality of the writing is good.
- Segmentation based sliding windows: [6] selects vertical strips of the image; this method can cause noise due to reconciliations adjacent characters.

B.Yanikoglu and P.Sandon [7] proposed another type of segmentation: segmentation using histograms projections. It based on the computation of projection histograms in several directions. This technique requires regular spacing in the image.

R.Manmatha et al [8] proposed words segmentation in scanned Latin documents using scale space technique. This method studied the behavior of the scale space to merge blobs of the words in the detected lines. However, the words segmentation can be achieved with a faster method. The mathematical morphology offers an effective way for merge the characters in this words.

Our method also based on the theory of scale space whose purpose is to break the bonds of pseudo-words found in images containing Arabic text.

3. PRINCIPLE OF THE METHOD

The segmentation method proposed in Arabic printed characters based on operations illustrated by the block diagram in the figure below:

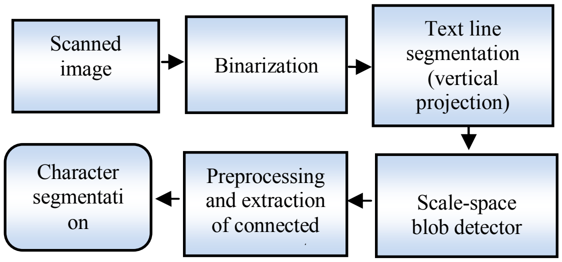

Scanned image
Binarization
Text line segmentation (vertical projection)
Scale-space blob detector
Preprocessing and extraction of connected
Character segmentati on

Figure 1: Block Diagram Of The Character Segmentation

3.1. Binarization

This binarization done through the application of a global thresholding method of the raw image obtained after acquisition. We opted for the method of global thresholding for various processed images that have a bimodal histogram expressing both classes of pixels of the image background (paper texture) and pixel content (diagrams, characters and all the spots ink). In this work, we used the global optimal thresholding method of Otsu [9].

The binary image produced by binary segmentation often contains a noise whose removal effected by the application of closing binary morphological filtering [10].

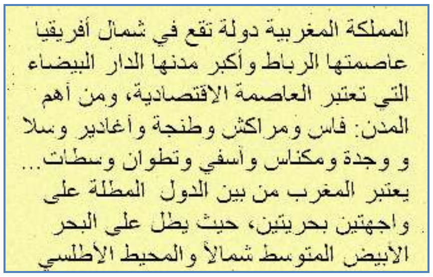

المملكة المغربية دولة تقع في شمال أفريقيا
عاصمتها الرباط وأكبر مدنها الدار البيضاء
التي تعتبر العاصمة الاقتصادية، ومن أهم
المدن: فاس ومراكش وطنجة وأغادير وسلا
و وجدة ومكناس وآسفي وتطوان وسطات...
يعتبر المغرب من بين الدول المطلة على
واجهتين بحريتين، حيث يطل على البحر
الأبيض المتوسط شمالاً والمحيط الأطلسي

Figure 2: Original Image

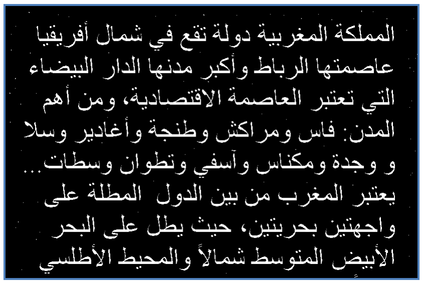
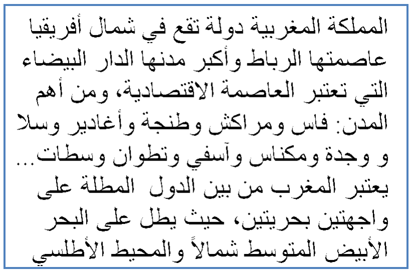

المملكة المغربية دولة تقع في شمال أفريقيا
عاصمتها الرباط وأكبر مدنها الدار البيضاء
التي تعتبر العاصمة الاقتصادية، ومن أهم
المدن: فاس ومراكش وطنجة وأغادير وسلا
و وجدة ومكناس وآسفي وتطوان وسطات...
يعتبر المغرب من بين الدول المطلة على
واجهتين بحريتين، حيث يطل على البحر
الأبيض المتوسط شمالاً والمحيط الأطلسي

Figure 3: Binarization Result



المملكة المغربية دولة تقع في شمال أفريقيا
عاصمتها الرباط وأكبر مدنها الدار البيضاء
التي تعتبر العاصمة الاقتصادية، ومن أهم
المدن: فاس ومراكش وطنجة وأغادير وسلا
و وجدة ومكناس وآسفي وتطوان وسطات...
يعتبر المغرب من بين الدول المطلة على
واجهتين بحريتين، حيث يطل على البحر
الأبيض المتوسط شمالاً والمحيط الأطلسي

Figure 4: Result Of Morphological Filtering (Closing)

3.2. Text line segmentation

The method used for segmentation is the vertical projection lines of binary images. This method is effective if the lines of text are not overlapping. Moreover, it has the advantage of being very fast. In the case of images of printed text, the lines are generally spaced enough for not creating overlap, making segmentation of lines rather delicate. In this case, it is common to use the method of the vertical projection. This method proceeds by first calculating the profile of the vertical projection of the filtered image.

$$f(y) = \sum_{x=0}^{w} I(x,y) \tag{1}$$

The profile of the vertical projection defined as follows:

Where the function f (y) is the vertical projection of the image I with the intensity value of a pixel (x,y) in the axis x. Given that y varies from 0 to h (height of the image) and x ranges from 0 to w (width of the image).

Figure 5 illustrates the vertical projection profile calculated for the image of figure 4. It provides various features on the main lines on the background of the binary image and the row heights.

The separation zones between the lines corresponding to the detected vertical projection profile of the trays. The width of the tray defines the separating strip between two lines. The minima correspond to the observed baseline.

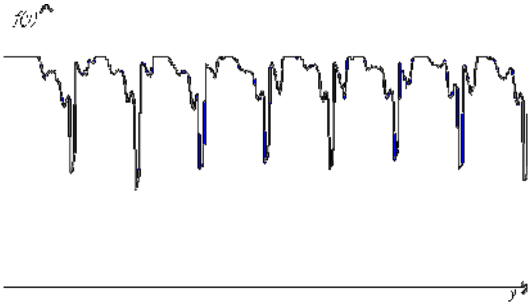
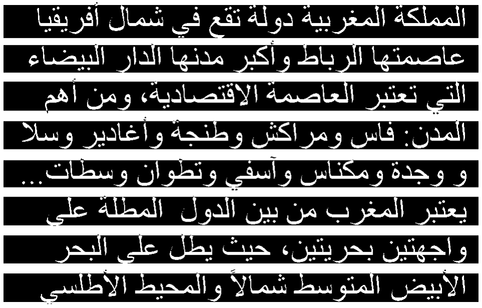


Figure 5: Vertical Projection Profile

Figure 6 illustrates the result of the method used for segmentation lines.



المملكة المغربية دولة تقع في شمال أفريقيا
عاصمتها الرباط وأكبر مدنها الدار البيضاء
التي تعتبر العاصمة الاقتصادية، ومن أهم
المدن: فاس ومراكش وطنجة وأغادير وسلا
و وجدة ومكناس وآسفي وتطوان وسطات...
يعتبر المغرب من بين الدول المطلة على
واجهتين بحريتين، حيث يطل على البحر
الأبيض المتوسط شمالاً والمحيط الأطلسي

Figure 6: Results Of Text-Line Segmentation

3.3. Scale-space blob detector

We adopted the theory of scale space for the selection of blobs in grayscale images containing printed Arabic then thresholding grayscale text. The main advantage of this selection is the character segmentation.

The theory of scale space was formalized by T.Linderberg [11]. This is to simulate the operation of the human eye when looking at a picture by placing it further from there it appears fuzzy. This appearance represented by smoothing the image using a Gaussian filter at scale σ.

The Gaussian scale space of an image I (x, y) defined by the convolution of the intensity of the image with a Gaussian filter.

A Gaussian filter G (x, y, σ) is defined by

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}) \tag{2}$$

I (x, y) is the image intensity function.

Calculating the convolution of the image I with a Gaussian filter at scale σ :

$$I(x,y)*G(x,y,\sigma)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}G(x-u,y-v,\sigma)I(u,v)dudv$$

$$I(x,y,\sigma)=I(x,y)*G(x,y,\sigma)$$

- Selection of circular Gaussian blobs

Circular blobs are blurs that observed on filtered image by a Gaussian at scale σ. The Laplacian of Gaussian LOG used to detect blobs [13]. The Laplacian of Gaussian at scale σ defined by

$$LOG(x,y,\sigma\ )=I_{xx}(x,y,\sigma)+I_{yy}(x,y,\sigma) \qquad (3)$$

We note that this type of blobs cannot characterize the Arabic printed characters. However, they promote contacts between characters and bonds.

- Selection of elliptical Gaussian blobs

An elliptical blob defined by the second derivative of the intensity in both σx scales along the x-axis, and σy along the y-axis. The elliptical Gaussian filter defined by:

$$G(x,y,\sigma_x,\sigma_y)=\frac{1}{2\pi\sigma_x\sigma_y}\exp(-\frac{1}{2}\left[\frac{x^2}{{\sigma_x}^2}+\frac{y^2}{{\sigma_y}^2}\right]) \qquad (4)$$

The scale-space convolution of the image I with a Gaussian at scale σx along the x-axis and σy scale along the y-axis is defined by:

$$I(x,y,\sigma_x,\sigma_y)=I(x,y)*G(x,y,\sigma_x,\sigma_y)$$

The Laplacian of Gaussian (LOG) at scale σx along the x-axis and scale σy along the y-axis defined by

$$LOG\,(x,y,\sigma_x,\sigma_y)=I_{xx}\,(x,y,\sigma_x)+I_{yy}\,(x,y,\sigma_y) \qquad (5)$$

The principle of the proposed selection of elliptical blobs defining Arabic printed characters method based on the choice of scales σx and σy. The second derivative of the Gaussian convolution kernel for σx and σy scales used to promote the appearance of elliptical blobs corresponding to variations of the light intensities in the direction of the y-axis (σy sufficiently large). However, the choice of σx can exponentially reduce the blobs corresponding to the connections of the characters (σx very small). After the selection of blobs phase, we perform thresholding to extract the binary masks of blobs forming the characters of Arabic printed text.

We propose an automatic choice σx and σy scales. This choice based on the selection of appropriate scales for the selection of blobs representing characters. The average value of heights of the segmented lines (Figure 5) is proportional to the scale σy. Therefore, the value scale σx can be deduced according to σy. After several tests σx values to locate the occurrence of blobs corresponding to characters, we set the value of the scale σx by

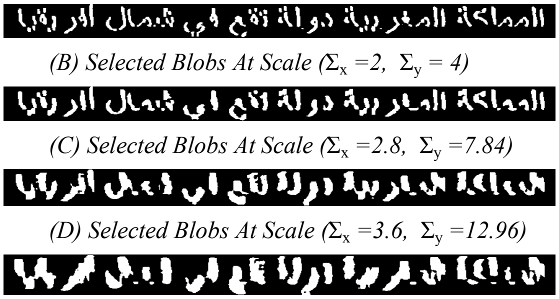
$$\sigma_x=\sqrt{\sigma_y} \qquad (6)$$

In this case, each line (Figure 6) of the text identified on the image. Each line contains words. Each word consists of pseudo-words, and isolated characters. The average height of the lines informs us about the automatic choice of scale space to use. For this reason, we have taken the first line segmented as an example. We illustrated five cases of convolutions of the original image with the Gaussian values (σx, σy) ={(1.2, 1.44) , (2, 4) , (2.8, 7.84) , (3.6, 12.96) , (4.4, 19.36) } with a pitch of 0.8 to scale σx. These five convolutions defined experimentally from tests on several documents containing different text sizes. The scale (σx, σy) = (4.4, 19.36) represents the biggest scale. However, the scale (σx, σy) = (1.2, 1.44) represents the lowest level. As many work that use the scale space Harris-Laplacian, Hessian-Laplacian [12] and SURF [13], we used as an initial low scale σx = 1.2 and a pitch equal to 0.8 between octaves.

We observed that the texture or the paper has very small changes in brightness. The sign of the Laplacian of Gaussian LOG is negative in this region, giving a black color for the background. Nevertheless, the sign of the Laplacian of Gaussian LOG is positive for important changes in the brightness of the characters. Writing thus represented by white blobs. The connecting lines of the characters interrupted for all the scales because they have small variations in the direction of the x-axis and characters show great variations in the direction of the y-axis. This is due to the choice of the scale σx (small value) and in the scale σy (high value). We note the appearance of blobs forming the characters that will be used as bit masks locating regions of interest. The following figure shows the appropriate scale for the correct detection of characters (Fig. 7 c).

المملكة المغربية دولة تقع في شمال أفريقيا

(A) Selected Blobs At Scale ($\Sigma_x = 1.2$, $\Sigma_y = 1.44$)



(B) Selected Blobs At Scale ($\Sigma_x = 2$, $\Sigma_y = 4$)

(C) Selected Blobs At Scale ($\Sigma_x = 2.8$, $\Sigma_y = 7.84$)

(D) Selected Blobs At Scale ($\Sigma_x = 3.6$, $\Sigma_y = 12.96$)

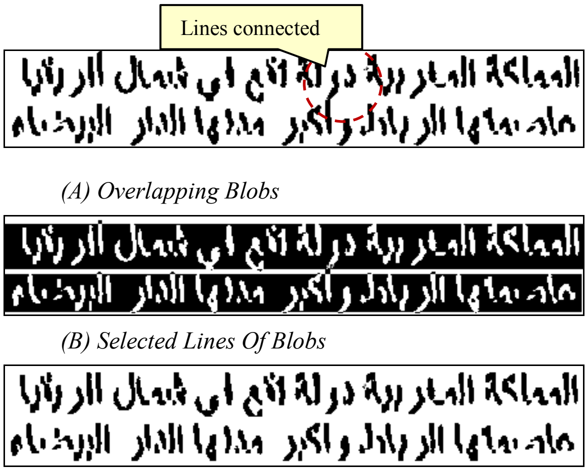
(D) Selected Blobs At Scale ($\Sigma_x = 4.4$, $\Sigma_y = 19.36$)

Figure 7: Results Of Selected Blobs In Characters At Different Scales

3.4. Preprocessing

In this section, we have solved the problem of overlapping lines caused by the Gaussian filter in the direction of the y-axis. We also solved the noise problem, which opposes the extraction of connected components of blobs characters. The additive noise is inherent in small size blobs caused by variations in brightness of the texture, or small black and white spots due to the binarization process. Several processes are necessary to restore the resulting image. To facilitate the task detection process areas of interest (connected components) and to help the segmentation of characters. We proposed the following solutions:

In the first solution, we solved the problem of overlapping blobs corresponding to the lines of text (Figure 8a) by applying the segmentation algorithm lines. Each line segmented already (Figure 6) is used as a mask selection of blobs lines (Figure 8b).



(A) Overlapping Blobs

(B) Selected Lines Of Blobs

(C) Result Of The Correction Of Overlapping Blobs

Figure 8: Correction Of Overlapping Lines At Scale (Σx =2, Σy = 4)

In the second solution, a noise can be characterized as a sudden change of a pixel insulated from its neighbors. We previously we recalled how filtered this noise in the binarization by a morphological filtering [10] (morphological closing of the image Figure 4). In this case, the morphological filtering can greatly affect the blobs characters. Therefore, we have proposed an alternative to locate the noise by performing a connected component labeling [14]. The objective is to enumerate the number of connected components by identifying the components of small areas that are not going to be consider in the next processing.

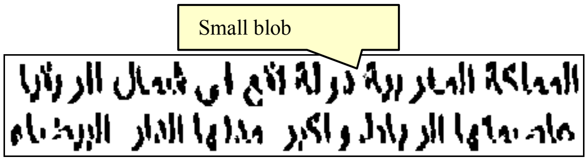


Figure 9: Localization Of Small Blobs

3.5. Character segmentation

The main goal in this step is to segment the characters. Each character enclosed in a rectangle as the case of figure 11. The boxed regions in rectangles are connected components of blobs on a suitable scale. The location result of blobs presented in figure 10. We applied the same detection algorithm of connected components [14] seen in the preprocessing phase. Small blobs are not taken into account. A small-connected component can be submerged with its neighbor to form the character of the text.

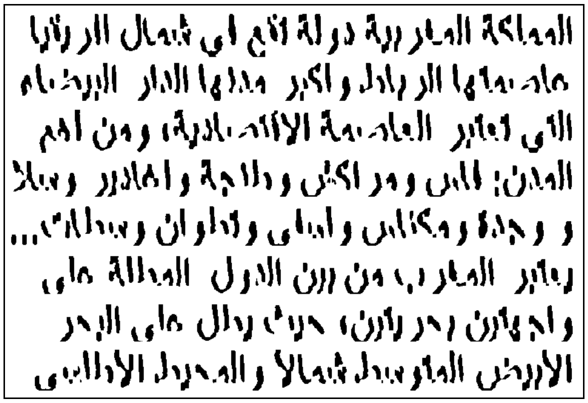
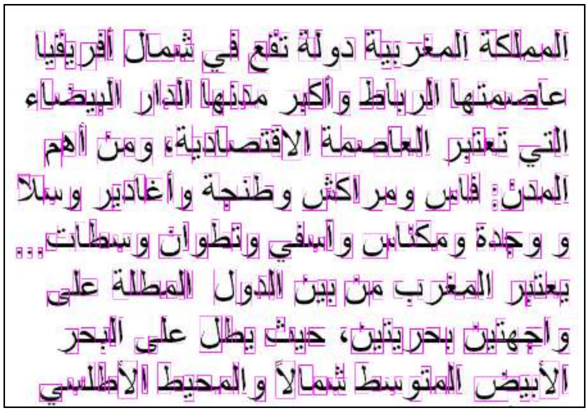


Figure 10: Blobs Detection Of Characters At Scale (Σx=2,Σy=4)



المملكة المغربية دولة تقع في شمال أفريقيا
عاصمتها الرباط وأكبر مدنها الدار البيضاء
التي تعتبر العاصمة الاقتصادية، ومن أهم
المدن: فاس ومراكش وطنجة وأغادير وسلا
و وجدة ومكناس وآسفي وتطوان وسطات...
يعتبر المغرب من بين الدول المطلة على
واجهتين بحريتين، حيث يطل على البحر
الأبيض المتوسط شمالاً والمحيط الأطلسي

Figure 11: Character Segmentation Results

4. DISCUSSION OF RESULTS

The Arabic alphabet is composed of more than twenty-nine characters (see table 1). Most letters slightly changing the shape of the character based on their position in the word (initial, medial and final). We have applied our segmentation method on dozens of images of Arabic printed text. The results are encouraging despite some problems of segmentation in the following letters: ش ، س ، ص ، ض، ط ، ظ. The character segmentation rates have identified for dozens of scanned images. We based our analysis on blobs that formed each character for calculate the segmentation rate. In these cases, the letter subdivided at most two or three fractions (see table 1).

Table 1: Arabic Alphabet And The Maximum Number Of Fractions Per Characters

Forms of letters				Name	Nbr max of fractions
Isolated	Final	Medial	Initial		
ا	ـا	ـاـ	ا	alif	1
ب	ـب	ـبـ	بـ	b	2
ت	ـت	ـتـ	تـ	t	1
ث	ـث	ـثـ	ثـ	tha	3
ج	ـج	ـجـ	جـ	j	1
ح	ـح	ـحـ	حـ	h	1
خ	ـخ	ـخـ	خـ	kha	1
د	ـد	ـدـ	د	d	2
ذ	ـذ	ـذـ	ذ	dhal	1
ر	ـر	-ـر-	ر	ra	1
ز	ـز	-ـز-	ز	z	1
س	ـس	ـسـ	سـ	s	3
ش	ـش	ـشـ	شـ	sh	3
ص	ـص	ـصـ	صـ	Ṣad	1
ض	ـض	ـضـ	ضـ	Ḍad	3
ط	ـط	ـطـ	طـ	Ṭa	2
ظ	ـظ	ـظـ	ظـ	Ẓa	2
ع	ـع	ـعـ	عـ	ayn	1
غ	ـغ	ـغـ	غـ	ghayn	1
ف	ـف	ـفـ	فـ	f	1
ق	ـق	ـقـ	قـ	qaf	1
ك	ـك	ـكـ	كـ	k	1
ل	ـل	ـلـ	لـ	l	1
م	ـم	ـمـ	مـ	m	1
ن	ـن	ـنـ	نـ	n	1
ه	ـه	ـهـ	هـ	ha	1
و	ـو	-ـو-	و	w	1
ي	ـي	ـيـ	يـ	y	1
ء	إ ؤ أ ئ			hamza	1
لا	ـلا	ـلاـ	لاـ	l+ alif	1

We have reached a rate neighboring 73% of good character segmentation. We did not exceed the rate of 10% for discharges of character segmentation. These releases can be improved with the addition of merging algorithm in the step of character recognition. The following table summarizes rates of Arabic printed character segmentation.

Table 2: Character Segmentation Rates For Dozens Of images

Max number of fractions	Number of characters	Segmentation rate
1	22	73.34 %
2	5	16.66 %
3	3	10 %

5. CONCLUSION

In this paper, we proposed an original method of Arabic character segmentation. We used the scale space in order to locate significant variations of intensity in different scales and select vertical elliptical blobs for each text character. We also used the vertical projection as extractor of lines. We exploited the connected components algorithm for the selection of blobs representing the characters.

The results are encouraging compared with other methods of segmentation of characters for the cursive text. Most of these methods using a sliding-window approach are very sensitive to noise and are not applicable in lower resolution documents. The degradation of the error rate recorded by our method is mainly due to the subdivision of some characters.

In perspective of future work, we believe to contribute to merge the fractions of each characters subdivided.

REFRENCES:

[1] T. lindeberg, Scale-space theory in computer vision Kluwer Academic Publishers, 1994.

[2] Wshah, S., Zhixin Shi, Govindaraju, V., "Segmentati on of Arabic Handwriting Based on both Contour and Skeleton Segmentation," 10th International Conf. on Document Analysis and Recogn. (ICDAR ‘09), 2009. pp. 793-797, 26-29 July 2009.

[3] "Segmentation and Recognition Strategy of Handwritten Connected Digits Based on the Oriented Sliding Window," in 2012 International Conference on Frontiers in Handwriting Recognition (ICFHR), Sept 2012, pp. 297–301

[4] X. Dupre. « Contributions à la reconnaissance de l'écriture cursive à l'aide de modèles de Markov cachés ». PhD thesis, Univ Rene Descartes - Paris V, 2003.

[5] Sriganesh Madhvanath, Venu Krpasundar, and Venu Govindaraju. Syntactic methodology of pruning large lexicons in cursive script recognition. Pattern Recognition, 34(1) :37–46, 2001.

[6] Y. Tay, P. Lallican, M. Khalid, C. Viard-Gaudin, and S. Knerr. An o□ine cursive handwritten word recognition system, 2001.

[7] B. Yanikoglu and P. Sandon. Segmentation of off-line cursive handwriting using linear programming, 1998.

[8] R. Manmatha, N.Srimal, Scale space technique for word segmentation in handwritten manuscripts, In SCALE-SPACE '99 International conference No2, Corfu, GRECE 1999, vol. 1682, pp. 22-33.

[9] N. Otsu, « A Threshold Selection Method from Gray-Level Histograms », IEEE transactions on Systems, Man and Cybernetics, 9(1), p. 62-66, 1979.

[10] J.Serra , Image Analysis and Mathematical Morphology, Academic Press, 1982.

[11] T. lindeberg, Feature detection with automatic scale selection, Technical report ISRN KTH NA/P--96/18--SE. Department of Numerical Analysis and Computing Science, Royal Institute of Technology, S-100 44 Stockholm, Sweden, May 1996.

[12] Bay H., T.Tuytelaars, L.V. Gool, « SURF : Speeded Up Robust Features », 9th European Conference on Computer Vision, Graz Austria, p. 404-417, May, 2006.

[13] Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In:ICCV. Volume 1. (2001) 525 – 531.

[14] JM Park, CG Looney, HC Chen, "Fast Connected Component Labeling Algorithm Using A Divide and Conquer Technique", Conference on computers and their Applications, 2000.