ISSN 2007-9737

# A Heuristic Approach to Detect and Localize Text in Arabic News Video

Sadek Mansouri[1], Mbarek Charhad[2], Mounir Zrigui[1]

[1] LATICE Laboratory, Research Department of Computer Science,
Tunisia

[2] Taibah University,
Saudi Arabia

mansouri_sadek@hotmail.fr, mbarek.charhad@gmail.com, mounir.zrigui@fsm.rnu.tn

**Abstract.** Automatic text detection in video sequences remains a challenging problem due to the variety of sizes, colors and the presence of complex background. In this paper, we attempt to solve this problem by proposing a robust detection-validation schema for text localization in Arabic news video. Candidate text regions are first detected by using a hybrid method which combines MSER detector and edge information. Then, these regions are grouped using morphological operators. Finally, a verification process is applied to remove noisy non-text regions including specific features for Arabic text. Performance and efficacy of the proposed text detection approach have been tested By using Arabic-Text-in-Video database (AcTiV-DB).

**Keywords.** Arabic text detection, LSD, AcTiV-DB.

## 1 Introduction

Every day, TV news provides a large number of video which allows people to follow the economic, political and social event in all the world. Moreover, thanks to the progress in mass storage technology, the amount of news video is growing rapidly especially on the World Wide Web (WWW). The diversity and the amount of these collections make access to useful information a complex task. Hence, there is a huge demand for efficient tools that enable users to find the required information in large news videos archives.

To achieve such an objective, several content-based video indexing systems have been proposed in the literature using two types of features: the perceptual features that are extracted in video frames such as color, texture, or shape. Although these features can easily be obtained, they do not give a precise idea of the image content. Extracting more semantic features and higher level entities have attracted more and more research interest recently. Text embedded especially the artificial text in video frames is one of the important semantic features of the video content analysis. This type of text is artificially added to the video at the time of editing and provides high-level information of the video content that seems to be a useful clue in the multimedia indexing system.

For example name of the speaker, headlines summarize the reports in the news video, place of the event and the score or name of the player (Figure 1). However, text detection and localization in a video frame is still a challenging problem due to the numerous difficulties resulting from a variety of text features (size, color and style), the presence of complex background and conditions of video acquisition. Although many methods have been proposed to detect embedded text, few methods are designed for text detection in Arabic news videos, due to Arabic language complexity and specific characteristics [8, 10, 7]. Indeed, the Arabic script is cursive and has various diacritical. An Arabic word is a sequence of related entities completely disjoint named Sub-words which are in turn formed of one or more characters.

Moreover, the complexity of the Arabic text is accentuated by the existence of dots, diacritics, stacked letters and different texture features

ISSN 2007-9737



**Fig. 1.** A examples of artificial text in news video

compared to Latin or Chinese. A specific feature of Arabic script is the presence of a baseline. Figure 2, demonstrates some of these characteristics on an Arabic text.

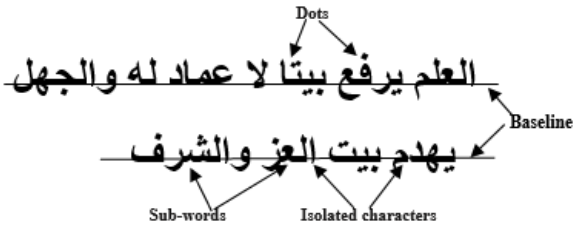


**Fig. 2.** Main characteristics of Arabic script

In this paper, we propose a novel method for Arabic text detection based on localization /validation schema. The localization process allows us to further extract candidate text region by using MSER detector and morphological operators. This detection process then avoids the processing on the entire image in order to have a faster algorithm. The second stage consists of two phases: the first is based on geometric of the detected regions, then we use specific features of Arabic text to refine detection and eliminate false detections. The proposed approach has been evaluated using the public AcTiV-DB dataset [20]. The rest of the paper is organized as follows: In Section 2, we discuss works related to text detection and localization. Section 3, presents our proposed approach and its different stages. Section 4, exposes experiments results and Section 5 states the conclusion.

## 2 Related Work

Many methods for text detection and localization have been proposed during the last few years based on different architectures, feature sets, and studies characteristics. These can generally be classified into three categories: connected component-based, edge-based, and texture-based.

The first category assumes that the text regions have a uniform color [9, 17]. In the first step, these methods perform in a color reduction and segmentation in some selected color channel as the red channel in color space as Lab space. Then they calculate the similarity of different color values to group neighboring pixels of similar colors into text region. Recently, Maximally Stable Extremal Regions (MSERs), based methods, have become the focus of several recent works for text detection [14], due to its robustness to scale changes and affinal transformation of images intensity. These approaches are based on the fact that the pixel intensity or color within a single text letter is uniform and they define the extremal region as a region which keeps stable in image binarization when modifying the threshold in a certain range. Experimental results show that MSER based method provides a high capability for detecting most text components. However, they also generate a large number of non-text regions, leading to high ambiguity between text and non-text in MSERs components.

The edge-based methods utilize some characteristics of text such as contrast of edge between texts, the background and the density in stroke to detect the boundaries of candidates text region. Then, non-text regions are removed by text verification process including some heuristic rules and geometric constraints. Anthimopoulos

et al. [3], use Canny edge detector to create an edge map. Then, they apply morphological operations (dilation and opening), on edges to construct candidates text region. Using some heuristics rules, an initial set of text regions is produced. In order to increase the precision of the detection rate, horizontal and vertical edge projections of each text region are performed. The authors in [15], use a sobel filter to detect contour on all frames of a video. Hence, they applied the morphological operations dilation and erosion to fuse the edges. Thereafter, same geometric constraints are selected to construct the coordinates of the text region.

The texture-base methods take into account the fact that text regions have special texture features different from another object of background. The first stage is to extract texture pattern of each block in the image by applying Fast Fourier Transform, Discrete Cosine transform, wavelet decomposition, and Gabor filter. Then a classification process is applied using k-means clustering, neural network and SVM in order to group each block into text and non-text region, et al. [16], proposed a novel method for image and video text detection. In the first step, a wavelet energy-based decomposition is performed. In the second pass, a texture features classification is applied in order to detect text lines using SVM for accurate text identification. Anthimopoulos et al. [4], proposed a two-stage schema for video text detection. Text line regions are firstly determined using edge detector and some heuristic rules. Then, obtained results are refined by an SVM classification based on edge Local Binary Patterns (eLBP). These methods face difficulties when the text is embedded in the complex background or touches other objects which have similar structural texture to texts.

Unlike Latin and English texts, few methods were designed to detect and extract the Arabic text from video sequences, some approaches have been proposed during the last years. Among them, we can cite the work of Ben Halima et al. [1]. In this work authors proposed a hybrid approach which combines colors and edges to detect Arabic text. Firstly, a multi-frame integration method was applied in order to minimize the variation of the image background.

Secondly, a set of features of colors and edges is used to localize the text areas. Alqutami et al. [2], use the laplacien operator to find the edge and k-means algorithm in order to classify all pixels into text region or non-text region. For text regions, they applied a projection profile analysis to determine the boundaries of text block. A similar approach was also presented by Moradi et al. [13], a sobel operator was used to extract edge. Then Morphological dilation was performed to connect the edges into clusters. Finally, a histogram analysis was examined to filter text areas. Sonia et al. [18], proposed three methods for Arabic text detection based on machine learning algorithms.

A Convolutional Neural Network was employed for extracting appropriate text image features and, clustering text, and non-text images. The other two proposed methods were based on multiexit boosting cascade. They learned to distinguish text and non-text areas using Multi-Block Local Binary Patterns (MBLBP), and Haar like features. In a more recent work, Oussama et al. [19], use SWT operator to extract connected component (CC), text candidates. CCs are filtered and grouped based on heuristic rules. Then Convolutional auto-encoders and the SVM classifier are applied in order to remove non text regions.

Our proposed approach differs from these approaches by employing a specific signature of Arabic text called baseline which may improve the detection task.

## 3 Proposed Approach

Our text detection method consists of two steps: text detection and text validation as shown in Figure 3. The first step detects connected components CC using a hybrid method which combines MSER and edge information. These CC are then grouped by mathematical morphology operators to construct candidate text regions. The second stage aims to remove non text region based on geometric constraints and a specific signature of Arabic script.

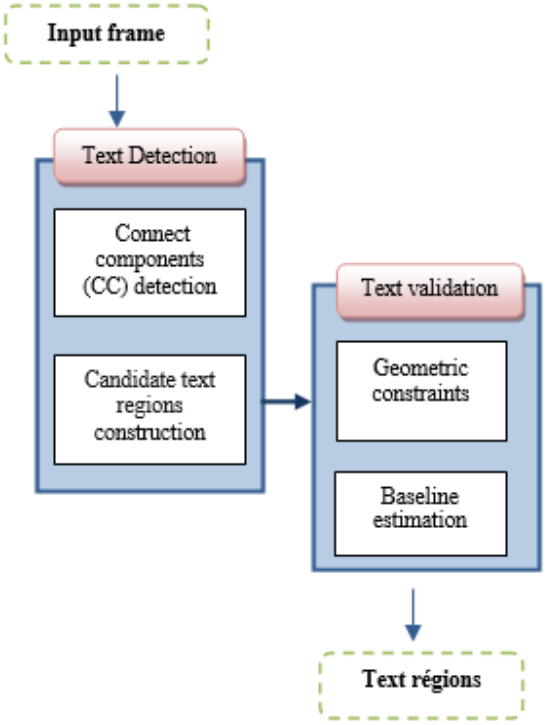


**Fig. 3.** Overall procedure of the proposed detection method

### 3.1 Text Detection

To be easily readable, the textual information embedded in video frames was written with distinct contrast to their background and in uniform intensity. Moreover, text lines produce strong edges horizontally aligned and follow specific shape restrictions. Our method made use of these basic features to detect text candidates regions in two steps.

#### 3.1.1 Connect Components (CC) Detection

Motivated by Chen work on Edge-enhanced MSER, we combine edge information and (Maximally stable extremal regions), MSER for extracting CC. In the first stage, MSER regions are efficiently extracted from the image. An ER (extremal region ), is a set of connected pixels in an image whose intensity values are higher than its outer boundary pixels Although MSERs are very stable and invariant to a new transformation of image intensities. The main problem arises from incorrect connections between noisy pixels and pixels of characters that provide a false detection for later operations. To overcome this shortcoming, an intersection operator between the canny edge and MSER regions is applied in order to remove incorrect connections between pixels.

#### 3.1.2 Candidate Text Regions Construction

The main goal of this step was to construct candidate text regions based on mathematical morphology operators. To do this, our proposed algorithm proceeded as follows: first off, we applied the closing operator in order to connect CC together. The closing of a gray-scale image $I(x, y)$, by a gray-scale structure element $B(s, t)$, was defined by:

$$F \bullet B = (F \Theta B) \diamond B, \tag{1}$$

where $F \Theta B$ denote the Dilation operator defined as follows: $F \Theta B(x, y) = maxF(x - s, y - t) + B(s, t)$. $F \diamond B$ denote the erosion operator: $F \diamond B(x, y) = minF(x + s, y + t) - B(s, t)$. Then, an open operator was applied aiming to reduce the small gaps and increase the holes between text blocks and other regions. Open of a gray-scale image $I(x, y)$, by a gray-scale structure element $B(s, t)$, was defined by:

$$F \circ B = (F \diamond B) \Theta B. \tag{2}$$

### 3.2 Text Validation

The results obtained (Figure 4), from the preceding stage constituted a first localization of the most likely text areas.

For the sake of reducing false detections, a filtering process seems to be an important task to validate candidate text regions. This filtering process was based on geometric properties and a specific signature of Arabic text called baseline.

Firstly, the obtained candidates regions (CR), are filtered according to their area as follows: $w * h \geq t_1$, where $w$ and $h$, denote respectively the width, the height of CR and $t_1$ refers to the fixed threshold to eliminate very small regions.

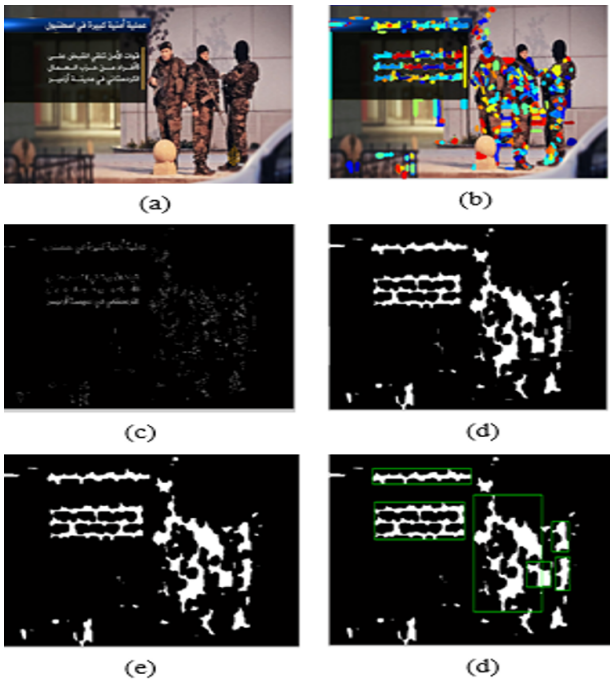


**Fig. 4.** Text detection: (a) original image, (b) MSER extraction in image (a), (c) Image mask integrating MSERs and canny edges (d) closing result, (e) open result (d) candidate text region

The second criterion was based on the fact that the artificial text is horizontally aligned. So, only the horizontal regions will be accepted according to the aspect ratio that should exceed an empirical predefined threshold Ts.

To make sure that the embedded text is included in the candidate region, we introduce new geometry descriptor called baseline which leads to the prominent feature of Arabic text. The baseline is a horizontal line whereas all word segments align and are connected to it, as shown in Figure 2. Major contributions have already been proposed in the field of printed and handwritten document. The horizontal projection is a common method founded on the fact that the words are horizontally aligned and separated by a similar distance between them. Consequently, the baseline will be determined according to the maximal peak in the pixels histogram. This Method work with binary image and can not automatically detect the baseline in video frames that have several challenges such as condition acquisition and complexity background.

The originality of our work is the proposition of a new method for baseline formation which is designed to improve the text detection in Arabic video frames. In a first stage, we use the fast Hough Transform to detect line segments for each word. the Hough transform algorithm uses a two-dimensional array, called an accumulator, to detect the existence of a line described by $r = x\cos\theta + y\sin\theta$. The dimension of the accumulator equals the number of unknown parameters, i.e., two, considering quantized values of r and $\theta$ in the pair (r,$\theta$). For each pixel at (x,y) and its neighborhood, the Hough transform algorithm determines if there is enough evidence of a straight line at that pixel. If so, it will calculate the parameters (r,$\theta$) of that line, and then look for the accumulator's bin that the parameters fall into, and increment the value of that bin. By finding the bins with the highest values, typically by looking for local maxima in the accumulator space, the most likely lines can be extracted.

As presented in Figure 5, The obtained line segments are likely to be fragmented and touch other non-text objects. To solve this problem, we propose to use some heuristic rules as follows:

**Rule 1:** Let consider N as the set of detected Lines segments in the image, a line segment where ($j \leq N$) is considered a candidate word segment if it meets the following conditions:

$$\Theta < \Delta\Theta, \tag{3}$$

$$L < \Delta l, \tag{4}$$

where $\Delta\Theta$ threshold over the direction and $\Delta l$ the maximal length of the segment that we should detect.

**Rule 2:** It is difficult to determine which line segment is a word segment based only on the length and orientation. More detailed information is required, including the distance between line segments. Based on the fact that Arabic words were horizontally aligned and separated by a similar distance between them, we define the regularity $R$ as a horizontal distance between

adjacent line segments. We compute R as in equation 4:

$$R = \frac{\sum_{i=1}^{N-1} |E_{i+1} - F_i|}{N-1}, \quad (5)$$

with $E$ and $F$ represent respectively the endpoint and first point of the line segment and N is the total number of line segments. A lines segment are considered as word segments, if the value of R is less or equals than a fixed threshold, otherwise these line segments is ignored. Since words segments have been successfully extracted, the last step consists of drawing the baseline passing through the first point of the first word segment and the endpoint of last word segment as illustrated in Figures.

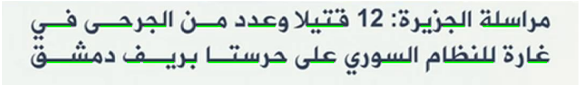


**Fig. 5.** Word segments detection results

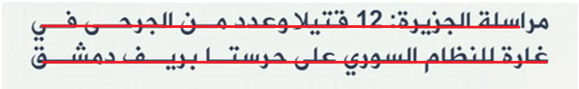


**Fig. 6.** Baseline estimation

The baseline detection has two main objectives: on one hand, to eliminate false alarms, such as candidate region which do not contain any textual information; and on the other hand, to refine text localization in candidate regions that contain text connected with some background items.

## 4 Experimental Results

### 4.1 Dataset

The proposed approach for Arabic text detection has been tested using the AcTiV-D [20], which containing 1843 frames distributed on two datasets.

#### 4.1.1 Dataset1 (high definition)

A set of video frames extracted from Aljazeera channel. This channel provides an image resolution (1920x1080), that is substantially higher than that of standard-definition.

#### 4.1.2 Dataset2 (standard definition)

A set of video frames collected from Al Wataniya 1, France 24 Arabic and Russia Today Arabic with low resolution (720x576).

### 4.2 Comparative Study for Baseline Estimation

To test the performance of our method for baseline estimation, we compare it to tow other methods in terms of precision and recall.

-HP: Horizontal projection method counts the number of black pixels having the same line position. The projection vector is represented by an histogram. In general, baseline position is determined by the central peak

-BLSD: In our pervious works [11, 12], we propose a new method for basline estimation called Blsd (bsaline estimation based on line segment detector). In a first step,line segment for each word has been detected using LSD method which is proposed by Grompone et al [6]. Then, we applied linear regression method to determine the parameters of the linear equation $y = mx + b$ of baseline.

**Table 1.** Performance comparison for baseline detection

| method | Precision | Recall |
|---|---|---|
| horizontal projection | 0.64 | 0.56 |
| BlSD | 0.76 | 0.83 |
| Our method | 0.85 | 0.92 |

The results show that our method based on hough transform present the highest presion and recall rates.

### 4.3 Performance of the Proposed System

In order to prove the effectiveness of the proposed method, a comparative study with previous systems is performed using precision, recall as the evaluation measures. We applied the evaluation method that has been proposed for the AcTiV-DB test set, together with evaluation results reported in [19], especially many-to-one matches method as shown in Figure 7.

The ground-truth is a set of rectangles $G = \{G_i, i = 1...n\}$. Indicating embedded text.

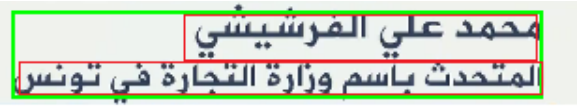


**Fig. 7.** Many to one match

**Table 2.** Experimental results of the text detection

| Channel | Method | Precision | Recall |
|---|---|---|---|
| HD(Aljazeera) | Chen [5] | 0.67 | 0.56 |
| | Zayene [19] | 0.85 | 0.83 |
| | our system | 0.90 | 0.87 |
| SD(france 24) | Chen [5] | 0.45 | 0.52 |
| | Zayene [19] | 0.75 | 0.73 |
| | our system | 0.71 | 0.70 |
| SD(RTArabic) | Chen [5] | 0.63 | 0.52 |
| | Zayene [19] | 0.73 | 0.73 |
| | our system | 0.75 | 0.74 |

Detection results consist in a set of rectangles $D = \{D_i, i = 1...m\}$, each containing detected texts. Recall and precision are computed as follow:

$$R = \Sigma_{i=1}^{n} \Sigma_{j=1}^{m} \left(\frac{w_R(G_i, D_j)}{n}\right), \quad (6)$$

$$P = \Sigma_{i=1}^{n} \Sigma_{j=1}^{m} \left(\frac{w_P(G_i, D_j)}{m}\right), \quad (7)$$

where

$$w_R(G_i, D_i) = \begin{cases} 1, if \ \ \forall i \in G, w_R(G_i, D_j) \geq t_r, \\ 0, otherwise, \end{cases}$$

$$w_P(G_i, D_i) = \begin{cases} 1, & if \sum_{i \in G} w_R(G_i, D_j) \geq t_p, \\ 0, & otherwise. \end{cases}$$

In our experiments, we used the area precision/recall thresholds proposed in the publication [19]: $tp = 0.4$ and $tr = 0.8$. Table I, show that the proposed system achieves excellent results for Alajazeera channel and it is able to outperform the method of Zayene [19]. We can notice the excellent precision rate of our method. This is due to the good rejection ability of false alarms using baseline desciptor. However, this higher score has been decreased For SD channels. Compared to Zayene work, we obtained the highest results for RT Arabic channel. But for France 24 Channel, Zayene work achieves the higher rate.


**Fig. 8.** Same results for our text detection method

As shown in Figure 8, the proposed method is robust to the variety of text size, style and color and complex background. We note that this work is designed only to detect static text, the dynamic text is out of the scope of this paper.

## 5 Conclusion and Future Work

In this paper, we have proposed an original approach for Arabic text detection in video frames. The main contribution includes the detailed analysis and the combination of text characteristics especially edge information and MSER regions to detect candidates text region. Moreover, our method for the baseline estimation considerably improves the filtering process allowing the system to remove the false detections.

Detailed experimental results and the comparisons with other methods are also reported

confirming that our proposed approach gives a good performance by the use of baseline feature. Hence its robustness to background complexity and text appearances are proven.

For future research, we will try to introduce the spatio-temporal information in our system aiming to detect embedded text in consecutive frames.

## References

1. **Alimi, A., Ben-Halima, M., & Karray, H. (2010).** Nf-savo: Neuro-fuzzy system-for Arabic video OCR. *International Journal of Advanced Computer Science and Applications*.
2. **Alqutami, A., Ahmad, A., & Atoum, J. (2011).** A robust algorithm for arabic video text detection. *Proceedings of International Congress on Computer Applications*.
3. **Anthimopoulos, M., Gatos, B., & Pratikakis, I. (2007).** Multiresolution text detection in video frames. *VISAPP*.
4. **Anthimopoulos, M., Gatos, B., & Pratikakis, I. (2010).** A two-stage scheme for text detection in video images. *Image and Vision Computing*.
5. **Chen, H. (2011).** Robust text detection in natural images with edge enhanced maximally stable extermal regions. *IEEE (ICPR)*.
6. **Grompone, R., Jakubowicz, J., Morel, J., & Randall, G. (2010).** LSD: A fast line segment detecto with a false detection control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
7. **Hkiri, E., Mallat, S., & Zrigui, M. (2016).** Events automatic extraction from Arabic texts. *IJIRR*, volume 6, pp. 36–51.
8. **Lhioui, C., Zouaghi, A., & Zrigui, M.**, **.** Realization of minimum discursive units segmentation of Arab oral utterances. *International Journal Computing Linguistics Appl*.
9. **Lim, J. & Park, J. (2007).** Text segmentation in color images using tensor voting. *Image and Vision Computing*.
10. **Mallat, S., Maraoui, M., Hkiri, E., & Zrigui, M. (2016).** Proposal of statistical method of semantic indexing for multilingual documents. *IEEE (FUZZ)*.
11. **Mansouri, S., Charhad, M., & Zrigui, M. (2017).** Arabic text detection in news video based on line segment detector. *Research in Computing Science*.
12. **Mansouri, S., Ihioui, C., Charhad, M., & Zrigui, M. (2017).** Text-to-concept: A semantic indexing framework for Arabic news videos. *18th International Conference CICLing*.
13. **Moradi, M., Mozaffari, S., & Orouji, A. (2010).** Farsi/Arabic text extraction from video images by corner detection. *Proceedings of 6th Iranian Conference on Machine Vision and Image Processing*.
14. **Neumann, L. & Matas, K. (2012).** Real-time scene text localization and recognition. *Proceedings of Conference on Computer Vision and Pattern Recognition*.
15. **Poignant, J., Besacier, L., & Quenot, G. (2012).** From text detection in videos to person identification. *IEEE International Conference on Multimedia and Expo (ICME)*.
16. **Ye, Q., Huang, Q., Gao, W., & Zhao, D. (2005).** Fast and robust text detection in images and video frames. *Image and Vision Computing*.
17. **Yi, C. & Tian, Y. (2011).** Text string detection from natural scenes by structure-based partition and grouping. *IEEE Transactions on Image Processing*.
18. **Yousfi, S., Berrani, A., & Garcia, C. (2014).** Deep learning and recurrent connectionist-based approaches for Arabic text recognition in videos. *Proceedings of The International Conference on Image Processing*.
19. **Zayene, O., Hennebert, J., Touj, S., Ingold, R., & Ben-Amara, N. (2016).** Text detection in Arabic news video based on SWT operator and convolutional auto-encoders. *Proceeding of 12th Workshop on Document Analysis Systems (IAPR)*.
20. **Zayene, O., Hennebert, J., Touj, S., Ingold, R., & BenAmara, N. (2015).** A dataset for arabic text detection, tracking and recognition in news videos – AcTiV. *Proceeding of ICDAR*, Nancy, France.

*Article received on 12/08/2016; accepted on 14/10/2016.*
*Corresponding author is Sadek Mansouri.*