ISSN 1870-4069

Arabic Text Detection in News Video Based on Line Segment Detector

Sadek Mansouri, Mbarek Charhad, Mounir Zrigui

LATICE Laboratry, Tunisia

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

Abstract. Text embedded in video sequences is very important to semantic indexing and content-based retrieval system, especially for large scale news collection. However, its detection and extraction is still an open problem due to the variety of its size and the complexity of the backgrounds. In this paper, we propose an approach for automatic Arabic-text localization based on a novel method for text-line detection. On the first stage, we use a line segment detector to detect candidate text lines. Then, we propose a word segment identification algorithm based on specific features for Arabic text in order to remove non-text lines. The last stage concerns the text line estimation and text detection in video frames. Experiment results, that we drove on a large collection of video images issued from news broadcasts show the excellent performance of our approach for text detection with different character sizes, directions and styles even in case of complex image background.

Keywords: Arabic text detection, line segment detection, baseline estimation.

1 Introduction

With the development of a big Arabic news channels, News video archives keep increasing in size every day and require more efficient tools for indexing and searching to facilitate access to these collections. Textual patterns embedded in video frames provide high-level information that seems to be a good way for semantic video annotation. However, its detection is still an open problem. This difficulty comes from the variation of style and size and complexity background.

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. In the first step, these methods perform in a color reduction and segmentation in some selected color channel as the red channel in or 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. Shivakumara et al. [1] extract connected components (CCs) using K-means clustering in the Fourier-Laplacian domain, and remove false detections using edge density, text straightness and proximity. Zhuge et al. [2] present a CC based algorithm which applied Maximally Stable Extremal Regions (MSER) as basic character candidates. Text CCs are then grouped into text regions using same geometric rules, and non-text regions are excluded based on corner detection and multi-frame verification.

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

The authors in [3] applied a Sobel filter to detect contour on video frames. Hence, they use the morphological operations to connect the edges together. Thereafter, the candidate's regions that respect the geometric constraints are selected to obtain the coordinates of the text boxes.

The method proposed by Yang et al. in [4] employ an edge based multi-scale text detector to extract text candidates that are then refined using an image entropy-based filter. Support Vector Machine (SVM) is applied as verification procedure to remove false alarms.

The texture-based methods take into account the fact that text regions have special texture features different from other object of background. The first stage is to extract texture pattern of each block in 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.

In [5] the authors propose a method using multi-oriented text detection which is based on the discontinuity of the text regions. To do this, they applied a Sobel mask and a Laplacian filter. Thereafter, Bayesian classifier is used to classify candidate pixels into text and non text regions. These methods face difficulties when the text is embedded in complex background or touches other objects which have similar structural texture to texts. Compared to Latin and Chinese text, few attempts have yet been designed to detect embedded text in Arabic news videos.

In this paper, we propose a novel approach for automatic Arabic text detection in news videos frames using a specific geometric feature of Arabic text called baseline in order to perform detection task. The baseline is defined as the imaginary line which connects all the characters of a word as shown in fig.1. Major contributions for baseline estimation have already been proposed in the field of printed and handwritten document. To the best of our knowledge, our approach is the first which use baseline for Arabic text detection in news video.

The remainder of this paper is structured 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 experiment results and section 5states the conclusion.

Fig. 1. Geometric features of Arabic text.

2 Related Works

Unlike Latin and English text, existing methods designed to detect and extract the Arabic text are very few, some approaches have been proposed during the last years.

Ben Halima and al. [6] propose a hybrid approach which combines color and edge to detect Arabic text. Firstly, a multi-frame integration method is applied in order to minimize the variation of the background of the image. Second, a set feature of color and edge is used to localize the text areas.

Alqutami et al. [7] use Laplacien operator to find an edge and k means algorithm to classify all pixels into text or not text region. For regions text, they apply a projection profile analysis to determine the boundary of text block. A similar approach was also presented by Moradi et al [8], a Sobel operator is used to extract edge. Then Morphological dilation is performed to connect the edges into clusters Finally, A histogram analysis is examined to filter text areas.

Sonia et al. [9] propose three methods for Arabic text detection based on machine learning algorithms. First, A Convolution Neural Network is employed for extracting appropriate text, image features and clustering text and non-text images. The two other proposed methods are based on multi-exit boosting cascade. They learn to distinguish text and non-text areas using Multi-Block Local Binary Patterns (MBLBP) and Haar-like features. The experimental results show that the neural network-based method outperforms the other proposed methods.

Recently, Oussama et al. [10] use SWT operator to extract connected component (CC) text candidate. The CCs are grouped based on heuristic rules. Then Convolution auto-encoders and 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 for embedded text detection in Arabic news videos frames.

3 Proposed System

The whole procedure of our text detection method is mainly divided into three stages as shown in fig.2. The first stage focuses on segmentation process based temporal information of video content. The second stage, presents the main contribution of our proposed approach including three steps: line segment detection, words segment selection, baseline estimation and text localization. In first step text lines and non-text lines in a video frame are identified based on line segment detection [11] algorithm. Second, we apply same heuristic rules based on geometric proprieties of Arabic words in order to remove false detections of text lines. Then, a linear regression method is used to estimate baseline and localize text region. Refinement stage aims to remove false detections.

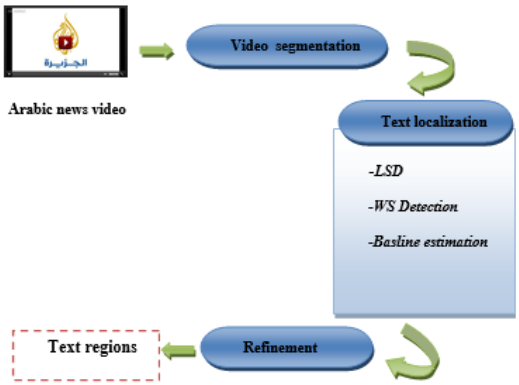

Fig. 2. Global overview of the proposed approach.

3.1 Video Segmentation

To analyze and understand its contents, the video needs to be parsed into segments. Most existing video database systems start with temporal segmentation of video into a hierarchical model of frames, shots and scenes. In this work, we applied a temporal segmentation based on the following assumption the text in the image requires at least two seconds being readable by the user, to generate shots. Then for each video shot, the middle image is selected as a key-frame.

3.2 Text Localization

Starting with the fact that the Arabic text is cursive, we define baseline as a set of word segments described by lines segments. The first step consists of detecting line segment in each key-frame.

Line Segment Detection To do this, we use the line segment detection algorithm that is proposed by Grompone Von Gioi [12]. This algorithm defines a line segment as a region called line-support region and it is based on the information of the gradient direction. Starting with the gradient image, we take a pixel which has a higher gradient magnitude as seed point. Then each adjacent pixel (Ap) which verifies the condition (1) will be added into line-support region and region angle is updated as formula:

$$abs(ang(Ap)le - \theta region) < \tau, \tag{1}$$

$$\theta region = arctan(\frac{\sum_i sin(ang_i)}{\sum_i cos(ang_i)}). \tag{2}$$

This step is repeated until no new pixel can be added to the line-support region. Then, the rectangular approximation of every line-support regions was extracted in order to determine the line segment. Each rectangle was defined by its center, length, width and orientation. In Line segment detection algorithm, the centroid of mass of the rectangular approximate is chosen as the center and the first inertia axis is used to determine the orientation of rectangle. The length and width are chosen in the way that covers the line-support region. Finally, to validate each line segment, a confidence index is calculated based on a contrario model that is proposed by [13].

Word Segment Detection Line segment detection algorithm provides excellent results for line segment detection as shown in Fig2 (b). However, 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 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 Δl 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 relationship between line segments. To this end, we define two types of distance: horizontal distance between adjacent line segment (hd) and vertical distance between parallel line segments (vd). Each distance will be used to remove false detection of word segments. Since words segments have been successfully extracted, the last step consists to estimate the baseline using on linear regression method.

Estimating the baseline is a useful task for the reader as well as for Arabic text extraction and recognition. However, its detection is a challenging task due to the wide variety of text visibility, such as variations in font and style and different lighting conditions. Major contributions have already been proposed in the field of printed and handwritten document [14]. The vertical projection is a common method [15] which based on the fact that the word was horizontally aligned and separated by a similar distance between them. Consequently, the baseline is determined according to the maximal peak in the histogram of pixels. All the pervious approaches work with binary image and it cant automatically detect the baseline in video frames that have several challenges such as condition acquisition and complexity background.

In our approach, the baseline is determined based on linear regression method. Let consider c=$\{c_1, c_2, ..., c_n\}$ where $c_i = (x_i, y_i)$ represents the center of word segments within the same direction. The baseline equation is defined by y=ax+b, where:

$$a = \frac{\Sigma_i y_i(\Sigma_i x_i) - n\Sigma_i x_i y_i}{D}, \tag{5}$$

$$b = \frac{(\Sigma_i x_i y_i)\Sigma_i x_i - (\Sigma_i y_i)\Sigma_i x_i^2}{D}, \tag{6}$$

$$D = (\Sigma_i x_i)^2 - n\Sigma_i y_i^2. \tag{7}$$

According to baseline coordinates, the next step of our approach consists of localizing text regions in video frames and representing each region by rectangular bounding box as shown in Fig 3(d).

3.3 Refinement

At this stage, we design a set of heuristic rules based on statistical and geometric properties of text regions to filter out false detections. First of all, we remove candidates regions with very large and very small aspect ratio. Then, we discard the detected regions which are located at the border of the image such as logo of TV channel. We note that the dynamic text will not be taken in account in this work.

4 Experimental Setup

4.1 Corpus

The proposed approach for Arabic-text detection has been tested on a large collection of video news frames. These videos are selected form four Arabic news channel: Al Arabia, Aljazeera (QATAR), AL WATANIYA (TUNISIA), Al Mayadeen (LIBANON) and characterized by the diversity of text pattern such as font, style, position size and background complexity in order to evaluate the robustness of our approach. Our dataset consists of 4000 frames distributed on two sets:

Fig. 3. Steps of text detection: (a) original image, (b) CC extraction, (c) candidates text regions and (d) final detection.

Dataset 1(high definition) a set of 2000 frames extracted from Aljazeera and Al Arabia channels. These channels provide an image resolution that is substantially higher than that of standard-definition (SD).

Dataset 2(standard definition) a set of 2000 frames collected from Al Wataniya 1 and Al Maydeen channels with low resolution. As the evaluation measures we have used recall, accuracy and false alarm.
As the evaluation measures we have used recall, precision and false alarm.

Table 1. Evaluation for our text detection method.

Dataset	Method	Recall	Precision	False alarm
Dataset 1(HD)	Our approach	0.72	0.81	0.28
	Epshtein [16]	0.6	0.63	0.37
Dataset 2 (SD)	Our approach	0.63	0.76	0.37
	Epshtein [16]	0.5	0.59	0.41

4.2 Results

Table 1 shows the experimental results of our method that we drove on two types for dataset. According to these results, it is clear that the proposed approach achieves good results for text detection using Dataset1 (HD) because these types

of channels provide an excellent quality of graphic text. Moreover, our method outperforms the method proposed in [16].

Fig. 4 shows some examples of text localization in video frames. We note that detected regions text will be filtered according to the minimal number of word segment.

Fig. 4. Steps of text detection: (a) original image, (b) CC extraction, (c) candidates text regions and (d) final detection.

We note that our method face difficulties when other objects have similar geometric characteristic to Arabic text as shown in fig 5.

Fig. 5. Same false detections of our method.

5 Conclusion

In this paper, we have presented a novel approach for text localization in the Arabic news video. A specificity of our proposal is to use the geometric features for Arabic text such as word segments and baseline in order to enhance text detection in video frames. Experimental results have shown that the proposed method is able to detect embedded text with different text appearances and complex backgrounds in HD channels and achieved higher recall and precision score than SD channels.

In future work, we plan to use other visual features to enhance detection task especially for video frames with low resolutions.

References

1. P. Shivakumara, T. Q. Phan, C. L.Tan: New Fourier-Statistical Features in RGB Space for Video Text Detection. IEEE transactions on Circuits and Systems for Video technology (CSV), pp. 1520–1532 (2010)
2. Y. Z. Zhuge, H. C. Lu: Robust video text detection with morphological filtering enhanced MSER. Journal of Computer science and Technology, pp. 353–363 (2015)
3. Poignant, J., Besacier, L., Quenot, G., Thollard, F.: From Text Detection in Videos to Person Identification. In: IEEE International Conference on Multimedia and Expo (ICME) (2012)
4. H. Yang, B. Quehl, H. Sack: A Framework for Improved Video Text Detection and Recognition. International Journal of Multimedia Tools and Applications (MTAP), pp. 217–245 (2012)
5. Shivakumara, P., Sreedhar, R. P., Phan, T. Q., Lu, S., Tan, C. L.: Multioriented Video Scene Text Detection through Bayesian Classification and Boundary Growing. IEEE Transactions on Circuits and Systems for Video Technology (2012)
6. A.M. Alimi, M. Ben Halima, H. Karray, A. Fernández Vila: Nf-savo: Neuro-fuzzy system for Arabic video OCR. International Journal of Advanced Computer Science and Applications, vol. 3, no. 10, pp. 128–136 (2012)
7. A. Alqutami, A.M.A. Ahmad, J. Atoum: A robust algorithm for Arabic video text detection. In: Proceedings of International Congress on Computer Applications and Computational Science, Advances in intelligent and Soft Computing, Bali, Indonesia (2011)
8. S. Zhang, C. Zhu, J. K. O. Sin, P. K. T. Mok: A novel ultrathin elevated channel low-temperature poly-Si TFT. IEEE Electron Device Lett., vol. 20, pp. 569–571 (1999)
9. M. Moradi, S. Mozaffari, A.A. Orouji: Farsi/Arabic text extraction from video images by corner detection. In: Proceedings of 6th Iranian Conference on Machine Vision and Image Processing, Isfahan, Iran (2010)
10. Yousfi, S., Berrani, S. A., Garcia, C.: Deep Learning and Recurrent Connectionist-based Approaches for Arabic Text Recognition in Videos. ICIP (2014)
11. O. Zayene, J. Hennebert, S. M. Touj, R. Ingold, N. E. Ben Amara: Text Detection in Arabic news Video Based on SWT Operator and Convolutional Auto-encoders. In: Proc of 12th IAPR Workshop on Document Analysis Systems (2016)
12. R. Grompone Von Gioi, J. Jakubowicz, J.M. Morel, G. Randall: LSD: a fast line segment detector with a false detection control. IEEE Transactions on Pattern

Analysis and Machine Intelligence, vol. 32, no. 4, Article ID 4731268, pp. 722–732 (2010)
13. A. Desolneux, L. Moisan, J. Morel: Meaningful alignments. International Journal of Computer Vision, vol. 40, no. I, pp. 7–23 (2000)
14. Almuallim, H., Yamaguchi, S.: A method of recognition of Arabic cursive handwriting. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 9(5):715–722 (1987)
15. Abu-Ain, T., Sheikh Abdullah, S., Bataineh, B., Omar, K., Abu-Ein, A.: A Novel Baseline Detection Method of Handwritten Arabic-Script Documents Based on Sub-Words. Soft Computing Applications and Intelligent Systems, Communications in Computer and Information Science, 67–77 (2013)
16. B. Epshtein, E. Ofek, Y. Wexler: Detecting text in natural scenes with stroke width transform. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)