A New Method for Arabic Text Detection in Natural Scene Image Based on the Color Homogeneity

Houda Gaddour[1(✉)], Slim Kanoun[1], and Nicole Vincent[2]

[1] Miracl Laboratory, Sfax University, Sfax, Tunisia
houda.gaddour@yahoo.fr, slim.kanoun@gmail.com
[2] Lipade Laboratory, Paris Descartes University, Paris, France
Nicole.Vincent@mi.parisdescartes.fr

Abstract. Text detection in natural scene image is still open research topics. Particularly, for Arabic text, a very few studies have been proposed. In this paper, we propose a method for Arabic text detection in natural scene image based on the color homogeneity. Starting from the MSER idea and instead of relying on a range of unique thresholds we calculate a range of pairs of thresholds for each channel in the RGB space in order to generate a set of binary maps. Following extraction of connected components of each binary map we apply a first filtering according to a stability criterion of the written texts to extract candidate components regardless of the language. Then, through the characteristics of the Arabic script we make a second screening to found candidates to keep only those that define a text in the Arabic language.

Keywords: Natural scene image · Arabic text detection · Color homogeneity

1 Introduction

Text detection in natural scene images is an important research subject for content based image analysis field. In this framework, several research works are proposed in lasts years. Nevertheless, we can notice generally several restrictions have been made in the studies. In fact, the type of the text can influence the choice of the used method. Text can be incrusted in the image or natural text written on flat surfaces or on any surface in scene images. Furthermore, the alphabet used for the text gives different aspects that cannot be handled in the same way. Unlike artificial or encrusted text, the natural text can be small and sometimes not readable because it is not intended to be. Then, natural text is difficult to detect and less work has been made in this direction. In the following, we will focus our study on text detection in natural scene image. Existing methods for scene text detection can roughly be categorized into two approaches: approach based on regions segmentation and approach based on texture and learning text properties.

The first approach makes image segmentation into regions and groups them in regions of characters and words. It is based on the color characteristics or on the pixel gray level areas or on the high contrast with the background through the binary information or the properties of the region contours. In [1] three specific text characteristics are applied to generate three contour maps as the Canny filter. For each candidate text

A. Mansouri et al. (Eds.): ICISP 2016, LNCS 9680, pp. 127–136, 2016.
DOI: 10.1007/978-3-319-33618-3_14

boundary, one or more candidate characters are then segmented with a local threshold based on the neighboring pixels. In [2], Sobel filter is used to create a contour map combining four contour maps according to the four directions i.e. horizontal, vertical, oblique top right and top left oblique. A model of scale space with N-levels is constructed and spatial responses to the Laplace Gauss filters are computed to generate a set of text candidates based on the character stroke width. A distribution of the strongest responses from the space scale model is used to check whether a candidate is a text area or not.

In some works, the color information is the basis of its proposed method. In [3], after color constancy and a noise reduction stage, the output image is passed through a color quantization step made by the minimum-variance method proposed in [4]. Indeed, pixels are grouped on the difference basis between their values. N binary maps are generated to retrieve a set of connected components (CCs). These CCs are passed through an initial screening where the regions are analyzed on the basis of geometric properties. A second screening based on the characteristics of the HOG descriptor is achieved. Park et al. [5] consider that the texts are homogeneous in space and use a labeling process that divides roughly an image into multiple layers. Noise pixels are then removed thanks to a median filter. Finally, chromatic and achromatic components are separated by a K-means segmentation method [6]. MSER technical [7] is the most used technique for reliable extraction of CCs. Moreover, the MSER was used as the pretreatment step for detection methods in [8–11], where the authors showed that this method is able to detect text characters as homogeneous components. The Stroke Width Transform (SWT) was their filtering method.

The second approach is based on the assumption that text is characterized by a dense area. This can be equated with a more or less regular pattern texture that enables to distinguish text from background. These texture properties are characterized by techniques based on the Gabor filter [12], the spatial variance, the wavelet [13], the Fourier transform, etc. Yi describes a method in [14] to locate text regions. First, adjacent characters [14] are grouped as candidate patch images. Then, features are extracted using Haar gradient maps. In [15] a texture-based approach is applied based on two characteristics namely contrast and color homogeneity applied on a segmentation of textured pattern using the EM segmentation method.

Most of the proposed works are based on binary or grayscale images through morphological operators, a contour analysis, gradients or wavelets. However, in case of noisy or poorly contrasted images, a detection system will be less efficient. Indeed, the color helps in complex environments. Some works based on color assume that text characters are monochrome and propose color-based methods founded on the dominant color of text through a color quantized histogram. In this case, the text boxes are assumed to have the same color index [16, 17]. This method works well when the text characters are perfectly monochromatic. However, it is not reliable if the processed image has low contrast between the foreground and background. Often it is useful to examine more than a color space [18]. To be able to touch all zones of different intensities (light or dark), the MSERs based methods, have become the focus of several recent works [10, 19] in order to detect all components having a stable uniform color. In this sense we note that the detection of a homogeneous area is not suitable with the search for pixels

below or above a given threshold on the opposite it is more significant if the search is between two thresholds interval defining.

In the following, we will focus our study on the natural Arabic text in scene images. Besides the lack of work for the Arabic language has motivated this study. We exploit effectively the color information for the text regions identification in the presence of surrounding noise, complex backgrounds and lighting problems which can degrade the color contrast of the text relative to the background. We propose a method based on the color homogeneity of the text regions and on some Arabic language characteristics to detect the text boxes contained in natural scene images. We prove, through various experiments, the contribution of this method in detecting Arabic texts regions appearing in a scene image.

2 Proposed Method

Text in a scene image generally has a homogeneous color with respect to that of the background. We present here two main steps of the process. First, text candidate regions are extracted and selected in any language with readability criteria and then they are filtered according to the Arabic alphabet characteristics.

2.1 Text Candidates Extraction

In most studies, text candidates extraction is based on a single threshold of a gray level image either a global [20] or a local [21] threshold. However, the use of a single threshold binarization opposes the search for a uniform color. Also, we can see in the real scenes images, it is not easy to know the precise zone of the text color and the approach even can fall in case when the text is polychrome. In our case, we consider homogeneity property in a more strict sense. Indeed, a double dynamic thresholding extracts color uniformity, more accurately than light or dark. This makes more sense in a color context. For this, a range of threshold pairs is applied to the image.

Binary Maps Generation. To fix these pairs of thresholds, we start by clustering the colors in the image to generate the most representative colors in the image. The k-means algorithm classifies the color pixels and creates clusters. In the RGB color space, each channel R, G, or B varies from 0 to 255. To each channel corresponds a grayscale image I_c (c = R or G or B). For each image I_c we apply a k-means to pixel color for dividing the color space into N zones, each of which defines a set of colors considered as similar in the image.

- Input of k-means: As the results of k-means clustering algorithm is depending on the initialization of the class centers, we have chosen to fix them in a deterministic way. N initial centers are uniformly distributed as:

$$C_k = min + k * ((max - min)/(N)) \quad (1)$$

min and max respectively represents the minimum and maximum value of gray levels in the image I_c and k ∈ [0, *N*–1].

- Iterations: at each iteration, we evaluate the membership of each pixel to *N* clusters and we associate it with the class with minimum Euclidean distance. Thereafter new centers are calculated from the clusters as:

$$C_k = \sum(x_{jk})/p_k \tag{2}$$

x_{jk}: is the value of a pixel *j* associated with the class *k* the cardinal of which is p_k when $k \in [0, N–1]$ and $j \in [1, p_k]$.

- Output of k-means: final clusters are characterized by their centers, they form different colors defined by a value interval with extremity *S1* and *S2* that are used as a pair of thresholds. The thresholds are calculated as follows:

$$\left.\begin{array}{l} S1_k = \left\{\begin{array}{ll} min & \text{for } k = 0 \\ S2_{k-1} & \text{for } 1 <= k < N \end{array}\right. \\ S2_k = \left\{\begin{array}{ll} (C_k + C_{k+1})/2 & \text{for } 0 <= k < N-1 \\ max & \text{for } k = N-1 \end{array}\right. \end{array}\right\} k \in [0,\ N-1] \tag{3}$$

Therefore, the found thresholds divide the axis of the gray levels into N distinct intervals for which $S1_k$ and $S2_k$ are the two extremities of the k^{th} interval. Following a series of experiments made, we empirically have chosen a value of *N* equal to 8 for our experiments corresponding to the number of colors that the human eye can easily distinguish. Therefore, 8 pairs of thresholds to 8 intervals, so 8 binary maps are found after binarization with two thresholds.

The binary map of each Ik_c interval is generated by assigning the value '1' for the pixels belonging to the interval while fixing the remaining pixel values to '0'. This processing function is reflected as follows:

$$IK_c(x_i) = \left\{\begin{array}{l} 1 \text{ if } S1_k <= x_i <= S2_k \ (x_i \text{ is the pixel number of the image } I_c) \\ 0 \ \textit{Otherwise} \end{array}\right. \tag{4}$$

The above process is repeated for each image I_c linked to the three channels R, G and B. Finally, 3 *N* binary maps are generated. The Fig. 1 shows the output of this step on a real scene image with '$N = 8$'. We can notice that irrespective of its color in the image, the text must necessarily appear at least in one of 24 (8*3) binary maps which guarantees us both a good extractor of connected components and that all text zones in the picture will be detected.

We will later extract the connected components from each of binary map and filter them in several stages to retain only the text candidates. These steps will be explained below.

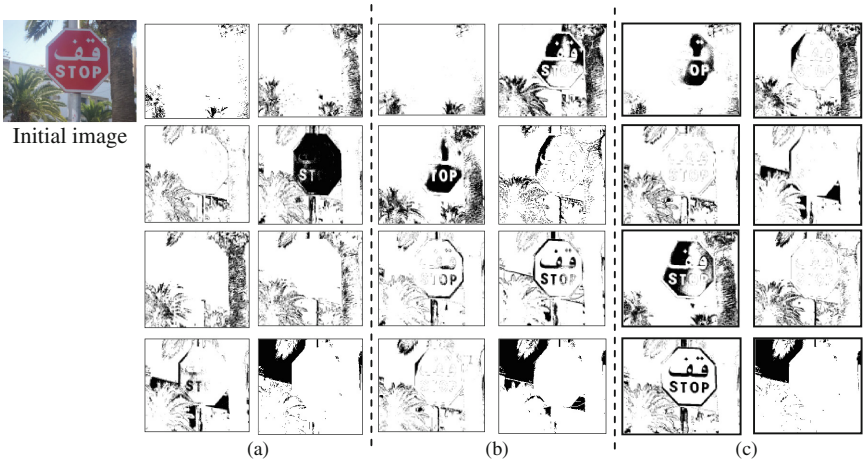

Fig. 1. All binary Maps for $N = 8$ (a) for I_R (b) for I_G (c) for I_B

Text Candidates Filtering According to Area Stability. In the binary maps we have defined the CCs that can be considered as text candidate regions. The MSER technique is based on the idea of taking the regions which remain nearly the same throughout a wide range of thresholds. We cannot use the same process because our CCs are not characterized in the same way, then we operate our first contribution, to have split color space into zones across a range of pairs thresholds, to test stability of candidate regions by varying the color areas intervals.

Indeed, we propose a new elimination criterion called double stability test. This criterion tests the stability of the regions through a range of thresholds pairs. In this sense, for each connected component CC_i looking at its evolution by varying the two extremities of the grayscale interval [*S1, S2*] either increasing (decreasing *S1* and increase *S2*) or reducing it (increasing *S1*and decreasing *S2*). For a text component, surface remains relatively stable since the writing is sharp and is contrasted with respect to the rest of the image. For a non-text component, in most cases there will be a greater variation in the surface of the component. This processing function is reflected as follows:

$$\text{Surface}(CC_i) - \text{Surface}(CC_{1i}) < \varepsilon \tag{5}$$

and

$$\text{Surface}(CC_i) - \text{Surface}(CC_{2i}) < \varepsilon \tag{6}$$

Where:

ε*:* is the area stability threshold.

$CC_{i:}$ is extracted from the binary image defined by two threshold [*S1, S2*].

CC_{1i}*:* is extracted from the binary image defined by two threshold [$S1 + y$, $S2 - y$] where y it's a level for reducing.

CC_{2i}: is extracted from the binary image defined by two threshold [$S1- y$, $S2 + y$] where y it's a level for increasing.

At the end of this step, we see in Fig. 3-b that all text written in any language was selected but also other non-text areas that we need to further filter according to specific characteristics for a particular language.

2.2 Arabic Text Candidates Extraction

Text candidates regions, extracted in the previous step, are first analyzed based on geometric properties (width, height, area). The regions that do not respect geometric lines are eliminated. The remaining candidates are mainly filtered with two Arabic features. Firstly, an entity in an Arabic text is usually a combination of several characters linked by ligatures to form a pseudo-word or a word. This is due to the cursive nature of the Arabic script. The ligature introduces a horizontal line between characters which has a stable height throughout the word. To exploit this property, we compute vertical projection profiles histogram of the CCs. On this histogram, ligatures appear with a constant height. Figure 2-a shows that the criterion of the existence of ligatures in Arabic words is very remarkable while for Latin words it does not appear. Secondly, the baseline corresponds to the line of pixels having the maximum value. If the word or pseudo-word has a single dominant peak in this case we can assume that it is an Arabic word. The vertical projection in Fig. 2-b illustrates the existence of the lonely peak that corresponds

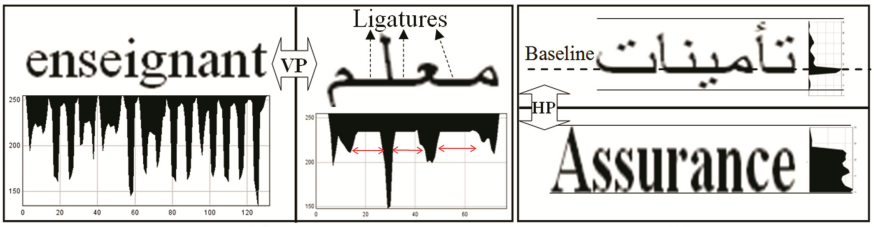

Fig. 2. Arabic text characteristics (a) Vertical projection (VP) on Latin and Arabic word (b) Horizontal projection (HP) on Latin and Arabic word.

Fig. 3. Text regions detected (a) All of the regions found in the extraction step, regions number = 329 (b) after filtering by color stability, text candidates number = 21 (c) after filtering according to Arabic features, Arabic text regions number = 1.

to the baseline in Arabic word. We note that for a Latin word, histogram is very different, but in Arabic word a second significant maximum is present.

Note that these two features are effective only for horizontal text regions. We notice in Fig. 3-c that we could eliminate false positive candidate region by such a process and only Arabic horizontal text regions are selected.

3 Experiments Results

The proposed method has been evaluated by using our dataset. Experimental results are then presented and discussed.

3.1 Evaluation of Arabic Text Extraction

As there is no public database of natural scene images designed for the detection of Arabic text, the proposed method was evaluated on our own database. To initiate it, we collected 50 images each containing written text areas together with the Latin and Arabic alphabet. All images in our database are captured natural scene images using a digital camera at 92 dpi resolution.

We manually built ground truth associated with this base in which we edited the coordinates of the bounding box of each pseudo-word in the image as well as its width and height. An entity is considered detected if its bounding box has a sufficient common surface in relation to that existing in the ground truth. The results for the experiments on text extraction are summarized in Table 1 where the number of existing pseudo-words, the number of detected pseudo-words, the number of false alarms and the corresponding values for recall and precision are listed.

Table 1. Experimental results for text extraction.

Number of images	50
Number of pseudo-words	307
Number of correct detected pseudo-words	274
Number of false positives detected pseudo-words	33
Recall rate	0.89
Precision rate	0.78

The algorithm of Arabic text detection gives a recall of 0.89 and an accuracy of 0.78. Thus, we conclude that our method achieves promising results. However, the accuracy is relatively low, which explains that the false positive rate is quite high. The detection of these non-text regions can be improved by adding other Arabic alphabet characteristics analysis.

3.2 Evaluation of Text Candidate Extraction

In order to show the contribution of our method we compare our text candidates' extraction algorithm based on a range of two thresholds to MSER algorithm based on a unique

thresholds range. To the output of the system, we applied the same stages of selections to generate Arab text candidates. The precision and recall of our system is relatively better with our proposed algorithm than with MSER algorithm for Arabic text candidates extracted as shown in Table 2.

Table 2. Comparaison of our text candidates extraction algorithm to MSER algorithm.

	with MSER algorithm	with our text candidates extraction algorithm
Recall	0.63	**0.89**
Precision	0.57	**0.78**

These encouraging results can be improved by incorporating other descriptors of Arabic texts to detect the variability of orientations and distortion of perspective to more robustly detect all type of text before being recognized and retrieved.

3.3 Discussion

The proposed method still has several limitations that appear either after the filtering step depending on color stability criterion or after the filtering step as characteristics of Arabic illustrated in the examples listed in Table 3 below. Several reasons are discussed in the following.

Table 3. Final results of detection on some other sample images.

Original image	All CCs	CCs detected after filtering by color stability	CCs detected after filtering according to the characteristics of the Arabic language
	 *(a)*	 *(b)*	 *(c)*
	 *(d)*	 *(e)*	 *(f)*

On the one hand, we notice the existence of false positives detected which has an influence on the accuracy rate. The detection of these non-text CCs can be improved by adding the analysis of other characteristics associated with the Arabic alphabet. In addition, some CCs texts, particularly Latin characters are not detected at the first screening (Table 3-b and e). For this, an improvement in the color quantization step is suggested either by using a three-dimensional k-means or a transition from RGB space to another color space such as HSV or L*a*b space for a good distribution of color levels and thereafter a range of more specific pairs of thresholds.

On the other hand, the results found for the image shown in Table 3-f shows that the proposed method is not able to detect the Arab texts oriented or curved. It seems it is required to set up a new descriptor for multi-oriented Arabic texts lines. Finally the isolated Arabic characters with no ligatures are not detected as letters ﻭ ,ﺍ, ﺯ us is shown in Table 3-c. They may be reintroduced at a later stage.

4 Conclusion and Future Works

We presented a new method for detection and localization of Arabic text in natural scene images based on color. The basic idea of this approach is the consistency of the text color that distinguishes it from the foreground and from the other existing objects in the same image. For this, we propose to use a range of pairs of thresholds to construct a set of binary images rather than dividing the color space in dark and bright. Then, by analyzing the color stability of the connected components, the filtering on the connected components is done regardless to language text candidate regions. They form the entrance to a second filtering according to a criterion related to the Arabic alphabet and get out Arabic regions.

In the future, we aim at enhancing the candidate filtering part in order to be effective for all types of texts and images with complex background. For this, an improvement in the binary maps generation step is suggested either by using a three-dimensional k-means or transition from RGB space to another color space such as HSV or L*a*b space for a good distribution of color levels and thereafter a range of more precise pairs of thresholds. On the other hand, it is needed to set up a new descriptor for multi-oriented Arabic texts lines.

References

1. Shijian, L., Chen, T., Tian, S., Lim, J.H., Tan, C.L.: Scene text extraction based on edges and support vector regression. Int. J. Doc. Anal. Recogn. (IJDAR) **18**(2), 125–135 (2015)
2. Sun, Q., Lu, Y.: Text detection from natural scene images using scale space model. In: Zhang, W., Yang, X., Xu, Z., An, P., Liu, Q., Lu, Y. (eds.) IFTC 2012. CCIS, vol. 331, pp. 156–161. Springer, Heidelberg (2012)
3. Fraz, M., Sarfraz, S.: Exploiting color information for better scene text detection and recognition. Int. J. Doc. Anal. Recogn. **18**(2), 153–167 (2015)
4. Heckbert, P.S.: Color image quantization for frame buffer display. In: SIGGRAPH, pp. 297–307 (1982)

5. Park, J.-H., Yoon, H., Lee, G.-S.: Automatic segmentation of natural scene images based on chromatic and achromatic components. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2007. LNCS, vol. 4418, pp. 482–493. Springer, Heidelberg (2007)
6. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: BSMSP, pp. 281–297 (1967)
7. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: BMVC, pp. 1–10 (2002)
8. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: CVPR, pp. 2963–2970 (2010)
9. Chen, H., Tsai, S., Schroth, G., Chen.: Robust text detection in natural images with edge-enhanced maximally stable extremal region. In: ICIP, pp. 2609–2612 (2011)
10. Felhi, M., Bonnier N., Tabone, S.: A skeleton based descriptor for detecting text in real scene images. In: ICPR, pp. 282–285 (2012)
11. Xiaoming, H., Shen, T., Wang, R., Gao, C.: Text detection and recognition in natural scene images. In: ICEDlF, pp. 44–49 (2015)
12. Jain, A.K., Bhattacharjee, S.K.: Address block location on envelopes using gabor filters. Pattern Recogn. **25**, 1459–1477 (1992)
13. Mao, W., Chung, F., Lam, K., Siu, W.: Hybrid chinese/english text detection in images and video frames. In: ICPR, vol. 3, pp. 1015–1018 (2002)
14. Chucai, Y., Yingli, T.: Text extraction from scene images by character appearance and structure modeling. Comput. Vis. Image Underst. **117**, 182–194 (2012)
15. Anouel, H.: Detection and location text in natural scene images: application to the detection of moroccan number plates. Doctoral thesis. Mohammed V-Agdal University (2012)
16. Kim, S.K., Kim, D.W., Kim, H.J.: A recognition of vehicle license plate using a genetic algorithm based segmentation. In: ICIP, vol. 1, pp. 661–664 (1996)
17. Gllavata, J., Ewerth, R., Freisleben, B.: Text detection in images based on unsupervised classification of high-frequency wavelet coefficients. In: ICPR, vol. 1, pp. 425–428 (2004)
18. Li, H., Doermann, D., Kia, O.: Automatic text detection and tracking in digital video. IEEE Trans. Image Process. **9**, 147–156 (2000)
19. Yin, X.C., Yin, X., Huang, K., Hao, H.W.: Robust text detection in natural scene images. IEEE Trans. Pattern Anal. Mach. Intell. **36**(5), 970–983 (2014)
20. Zhong, Y., Karu, K., Jain, A.: Locating text in complex color images. In: ICDAR, vol. 1, pp. 146–149 (1995)
21. Ohya, J., Shio, A., Akamatsu, S.: Recognizing characters in scene images. IEEE Trans. Pattern Anal. Mach. Intell. **16**, 214–220 (1994)