A New Morphology-based Method for Text Detection in Image and Video

Zaynab El khattabi
Faculty of Sciences
LIROSA Laboratory-Tetouan, Morocco

Youness Tabii
National School of Applied Sciences,
LIROSA Laboratory -Tetouan, Morocco

Abdelhamid Benkaddour
Faculty of Sciences,
LIROSA Laboratory-Tetouan, Morocco

ABSTRACT

Text supply crucial suggestions for understanding video content, also the information that the text convey is much more concise than corresponding audio or video. The reason is that we need language knowledge to understand the text, and the knowledge itself does not need to be embedded in the text data. Text streams contain very rich semantic information. How to effectively extract information from text is an important component in video content analysis and semantics research. In this paper, a new morphology-based method for text detection in image and video is proposed. It consists of three major stages. In the first stage, the input color image is converted to gray-scale, a morphological binary map is generated by calculating the difference between the closing image and the opening image, and a binarization is performed. In the second stage, candidate regions are connected by using a morphological dilation and erosion operations. In the last stage, the extracted regions are verified based on characteristic of text regions to eliminate non text regions.

Keywords:

Information retrieval, Text detection, Morphology, Text recongnition

1. INTRODUCTION

Semantic content in images carries rich and important information which can be very useful in many areas of computer science, like content based image retrieval, multimedia databases indexing and classification purposes. This content can be objects, color, texture or shape as well as the relationships between them.
Textual information within an image or video is very interesting, because text can be used to describe the contents of image or video, and can be extracted and compared to other semantic contents. The two main tasks to extract text from image/video are: text detection and text recognition. The first step is the localization and detection process of the regions in image or video that contain texts, the second step is the recognition of characters in the regions extracted using the optical character recognition (OCR) algorithm to translate them into machine-encoded text. There are two types of texts in images or video, namely, the scene texts and the artificial texts. Scene text is textual content that was captured by the camera as part of a scene such as text on T-shirts or road signs [2]. Artificial text, as the name implies, is artificially added in order to describe the content of the video or the image or give additional information related to it. This makes it highly useful for building keyword indexes.
Although several techniques have been developed on text detection in real applications, fast and robust algorithms for detecting text from videos and natural scene images under various conditions need to be further investigated. Indeed, variations of text due to differences in size, style, orientation, color and alignment, as well as low image contrast and complex background make the problem of text detection and recognition from images or video one of the most challenging problems in Computer Vision.

This paper is organized as follows. In Section 2, a review of the related works on text detection and recognition in images is given. In Section 3, a text detection method based on morphological operations is presented. Experimental results and their discussions are presented in Section 4 and the conclusion is given in Section 5.

2. RELATED WORKS

There exist two main approaches related to the text detection problem: Region-based methods and texture-based methods.

Region-based methods use the properties of the color or gray scale in a text region or their differences with the corresponding properties of the background [6]. Generally, a region-based method consists of two stages: text detection to estimate text existing confidence in local image regions by classification, and text localization to cluster local text regions into text blocks, and text verification to remove non-text regions for further processing. These methods can be divided into two sub-approaches: connected component (CC)-based and edge-based.
Connected component (CC)-based: use bottom up approach to group smaller components into larger components until all regions are identified in the image. A geometrical analysis is needed to identify text components and group them to localize text regions. At first, certain local features are calculated for each pixel in the image and then pixels with similar feature values are grouped together using connected component analysis to form characters. The biggest drawback is sensitivity to noisy and low-resolution images, because they require low variance of local features in sufficient number of pixels [7].The basic steps of the connected component based methods are presented in [5]. In the first stage of preprocessing, the input image is converted to YUV space and converted to an edge image, which is sharpened in order to increase contrast between the detected edges and its background. In the process of localization, horizontal and vertical projection profiles are

computed of candidate text regions, which are segmented based on adaptive threshold values, calculated for the vertical and horizontal projections respectively. Only regions that fall within the threshold limits are considered as candidates for text. Rampurkar et al. [11] present a method for text detection in natural scene images based on a binarization and enhancement technique, followed by a CC analysis procedure to define the final binary images that consist of text regions. Yi-Feng et al.[8] present hybrid method to localize text in scene images. it consists of three stages: preprocessing by designing a text region detector to generate the text confidence map, based on which text components can be segmented by local binarization, CC analysis and text line grouping.

The performance of CC based methods is sensitive to text alignment orientation. Most of these methods cannot segment text components accurately without prior knowledge of text position and scale.

Edge-based: focus on high contrast between the background and text and edges of the text boundary.The basic steps of the edge-based text extraction algorithm are given in [3] as bellow: Creation of a Gaussian pyramid by filtering the input image with Gaussian kernel, down sample the image in each direction by half, convolve it with directional filter at different orientation kernels for edge detection, create a feature map associating a weight factor with each pixel to classify it as candidate or not for text region, carryout the dilation operation to enhance the text regions by eliminating non text regions and perform area based filtering to eliminate noise blobs. Rampurkar et al. [11] present an algorithm to detect horizontally aligned text in images and video based on the application of a color reduction technique and edge detection that focuses the attention to areas where text may occur by localization of text regions in the edge image using projection profile analyses and geometrical properties.

According to the results of the comparison between CC-based and edge-based methods in [3], both the approaches does not take into consideration the removal or noise or unwanted clutter from the test images before or after the computations. Therefore, a morphological cleaning operation can helps in reducing the number of false positives obtained and achieving a higher precision rate.

Texture-based methods use the observation that text in images has distinct textural properties that distinguish them from the background. The techniques based on Gabor Alters, Wavelet, FFT, spatial variance, etc. can be used to detect the textural properties of a text region in an image [6]. En general, text regions possess a special texture because text usually consists of character components which contrast the background.

Y. Zhong et al. [13] proposes a texture-based caption text localization method which operates directly in the DCT domain for MPEG video or JPEG images. Caption text regions are segmented from background images using their distinguishing texture characteristics and each unit block in the compressed images is classified as either text or non text based on local horizontal and vertical intensity variations encoded in the DCT domain. In addition, post-processing procedures including morphological operations and connected component analysis are performed to refine the detected text. S. Angadi et al. [1] present a texture based method in the DCT domain using a high pass filter to remove constant background. The contrast and homogeneity features are obtained on every block of the processed image using a gray level co-occurrence matrix, text blocks are identified with discriminant functions. In [9], a hybrid approach is proposed of text segmentation using edge and texture feature information. The Edge detection is performed and the edge, homogeneity and contrast features are calculated for each block.Then, blocks are assigned as text and non text on the basis of these features and are merged to obtain text regions.

Since texture based methods decrease the dependency on the text size, they have difficulty to find accurate boundaries of text areas [4]. All these works make evident that the text regions cannot be perfectly extracted from the image, because images consist of complex backgrounds and objects. As a result, it is common to obtain false text detections and misses.

3. TEXT DETECTION AND EXTRACTION

In this section, the proposed method of text localization and extraction from image and video with simple and complex background and texts of different size, font and alignment is presented. This approach is morphology-based, which relies on the fact that text regions present a special kind of features that makes them different from other image regions. It consists of three major stages. First, a preprocessing of the image is performed to generate a morphological binary map. Then, candidate regions within the binary map are connected by using two basic morphological operations. Finally, text region verification is necessary to discard non text regions. Figure 1 shows the flowchart of the proposed method.

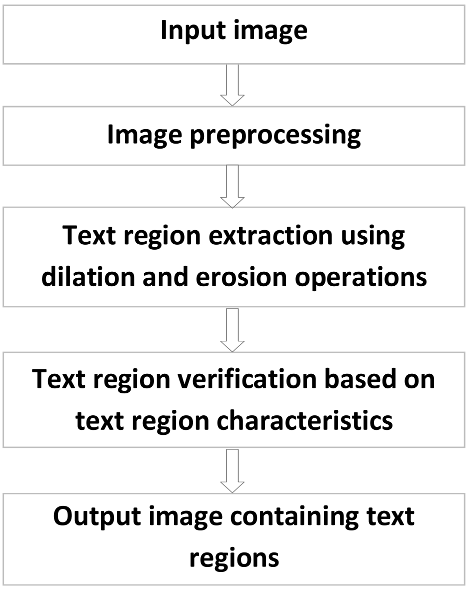
Input image
Image preprocessing
Text region extraction using dilation and erosion operations
Text region verification based on text region characteristics
Output image containing text regions

Fig. 1. Stages of the proposed method

3.1 Image preprocessing

In case of RGB space, the image will be transformed to gray scale space using the luminance Y component of YIQ color model. Equation (1) gives the conversion formula.

$$\begin{bmatrix} Y \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \qquad (1)$$

The resulting gray-scale image has only luminosity/brightness attribute which make easier to work on it. Then, a morphological binary map is generated, it is described in [10] with our modifications by using the following steps:

—Step 1: Generate two images (closing image and opening image), by applying morphological closing operation and opening operation to the input image with a disk structural element S_5.

—Step 2: Calculate the difference between the closing image and the opening image to increase the contrast between the possibly interesting regions and the rest of the image. Usually, there is a contrast between characters of text region and its background.

—Step 3: Apply a low pass filter with a 5x5 mask to smooth out noises and boundaries between text and background. In ideal cases, the pixel intensities within the same category (background or text) are equal, Intensities differences only occur at the boundary between text and background.

—Step 4: Perform Otsu image binarization, using a threshold defined dynamically according to the background of the image.

—Step 5: The noise introduced in the binarization stage is removed from the morphological binary map (step 3), by applying a low pass filter.

3.2 Candidate text region extraction

In order to connect candidate regions and reduce noises, morphological operations is used to the morphological binary map. Dilation is an operation which enhances the region of interest, using a structural element of the required shape and or size. The process of dilation is carried out in order to enhance the regions which lie close to each other. Following to this process,an erosion of the dilated image is performed to remove small spurious bright spots in the image. Let *M(x,y)* denote the morphological binary map and S_r denote a disk structuring element. Besides, let $\oplus$ denote a dilation operation, and □ denote an erosion operation. The process of dilation and erosion is described in algorithm 1.

Algorithm 1:Process of dilation and erosion

```
Input: morphological binary map M(x,y).
Output: resulting image with text candidate regions E(x,y).

Begin
For i:=1 to 3 step 1 do
D(x,y) := M(x,y) $\oplus$ $S_3$;
M(x,y) := D(x,y) ;
End

For j: =1 to 2 step 1 do
E(x,y) := D(x,y) □ $S_2$;
D(x,y) := E(x,y);
End
End
```

As a result of this step, the small regions are combined with left or right nearest region. After all text candidates regions are localized, regions extraction is getting from the gray-scale image by calculating the difference between the binary image and the gray-scale input image. The result of this stage is shown in Figure.2.

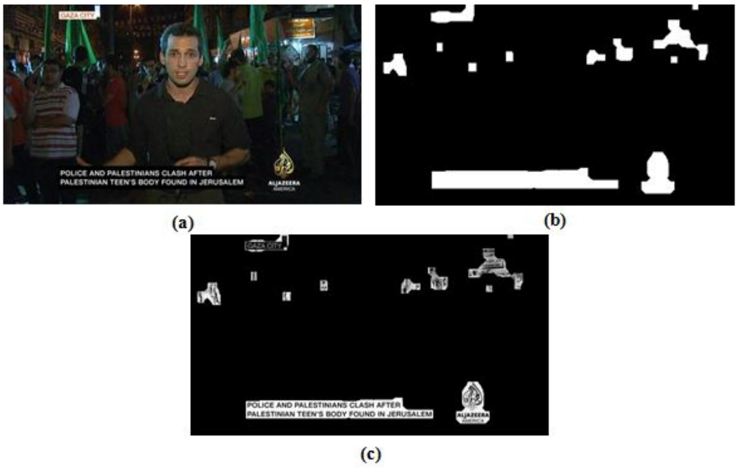
GAZA CITY
POLICE AND PALESTINIANS CLASH AFTER
PALESTINIAN TEEN'S BODY FOUND IN JERUSALEM
ALJAZEERA
(a)
(b)
(c)

Fig. 2. Candidate text region extraction. (a) Input Image. (b) Morphological binary map after dilation and erosion operations. (c) Region extraction from the gray scale image.

3.3 Text region verification

Several rules are used to filter out the false detections (i.e. noise regions and non text regions). The rules are based on characteristic of text regions. After decomposing the image into a set of regions, a verification of the extracted regions based on characteristic of text regions is performed to eliminate non text regions. Kaihua et al. [14] propose 12 features to expose the intrinsic characteristics of text connected components. In this stage, some of these features are used to remove the false detections.

—The first feature is the size of region, if a region is too small, it is discarded.

—The second feature used in [12], is the Aspect ratio of the region. If the value is too large or too small, the region is discarded. it is defined as:

$$AspectRatio = \frac{Height(region)}{Width(region)} \tag{2}$$

Long bar-shaped regions and too thin regions can be removed by this rule.

—Two shape regularity features are used from [14]; the occupancy ratio and the compactness ratio defined as:

$$OccupancyRatio = \frac{Area(region)}{Area(BoundingBox(region))} \tag{3}$$

$$CompactnessRatio = \frac{Area(region)}{(Border(region))^2} \tag{4}$$

These features are used to suppress the non text regions which have irregular shape but have strong texture response. When the occupancy ratio and the compactness are too small the regions are eliminated.

Figure 3 shows the resulting image of this stage.

Before applying an OCR to recognize the text, the refined text regions need to be converted to a binary image, where all pixels belonging to text are highlighted and others suppressed. Therefore,

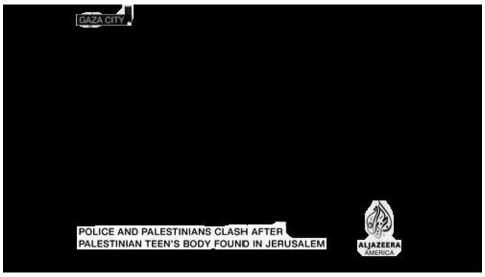
GAZA CITY
POLICE AND PALESTINIANS CLASH AFTER
PALESTINIAN TEEN'S BODY FOUND IN JERUSALEM
ALJAZEERA
AMERICA

Fig. 3. Resulting image after text region verification.

a binarization can be performed using a dynamic threshold which depends on the background of the image.

4. EXPERIMENTAL RESULTS

The proposed approach has been tested on scene images, video frames and images with artificial text embedded. A wide variety of text fonts, colors and orientation were represented. The proposed method has produced good results for video including television commercials and news broadcasts, where text regions are difficult to detect because some characters of them have different size and with complex backgrounds. In contrast to videos, this method shows efficient result in detecting scene and overlay text. Tested images are classified into 3 classes:
Class I: Scene images.
Class II: Video images.
Class III: Images with overlay text.

Figure 4 presents examples of text detection from images of the 3 classes. Regarding the detection speed, it was observed that the processing time lies in the range of 0 to 7 seconds due to varying resolution of image. Furthermore,the method is tested to detect Arabic language and it produced acceptable results, but it still needs improvements to increase the detection rate in other languages.

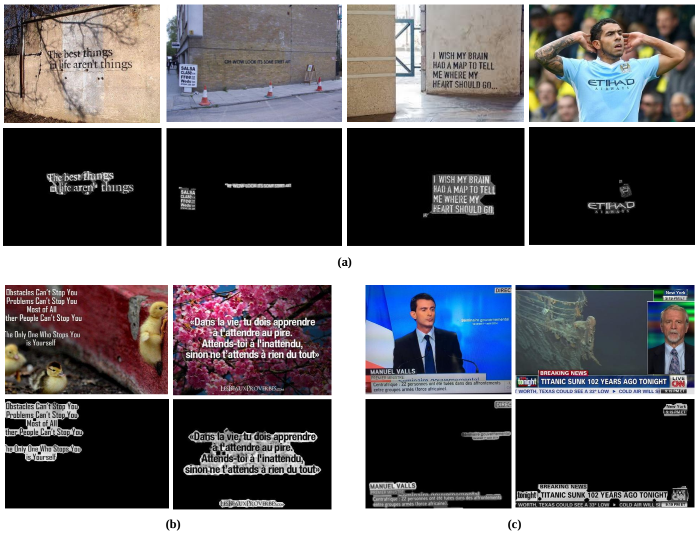

Fig. 4. Experimental results (a) Class I: scene images.(b) Class III: images with overlay text.(c) Class II: video images.

The detection performance of processing different types of images is given in Table 1.

Table 1. Text detection performance

	Class I	Class II	Class III
Total of images	30	40	20
Total of text regions	177	204	110
Detection Rate	91.52%	83.33%	92.72%
False Acceptance Rate	11.29%	4.9%	4.54%
False Rejection Rate	8.47%	16.66%	7.27%

Detection Rate = (Number of text regions correctly detected/ Number of text regions tested) * 100
False Acceptance Rate = (Number of text regions false detected/ Number of text regions tested) * 100
False Rejection Rate = (Number of missed text regions / Number of text regions tested) * 100

The method achieves an overall detection rate of 92.72%, and 91.52% for overlay and scene text. The reason for false rejection rate in scene images and images with overlay text, that are 8.47% and 7.27% successively, is the low contrast of text regions or parts of the text and the false acceptance rate of 11.29% in scene images is due to complex backgrounds containing trees, buildings and vehicles which are handled as text regions. However, In videos, the very large and small size of some characters which are difficult to be detected, is the reason of obtaining a false rejection rate of 16.66%.

5. CONCLUSION

This paper presents a new efficient morphology based approach for text detection. This method provides both rapidity and accuracy to extract text from different types of images and it is able to locate text regions with different sizes, orientations and different styles, even in case of texts occurring within complex image background. The proposed method consists of three major stages. Preprocessing of the input image is performed to generate a morphological binary map. Secondly, candidate regions are connected by using morphological operations. Finally, verification of the extracted regions based on characteristic of text regions is required to remove non text regions.

The effectuate experimentation on video and scene images shows that the approach is efficient and it can be extended to detect text in other languages. Moreover, in future works, temporal information can be used to increase the text detection rate in video, because temporal information may help in locating exact text position in video sequences.

6. REFERENCES

[1] S. A. Angadi and M. M. Kodabagi. A texture based methodology for text region extraction from low resolution natural scene images. In *Advance Computing Conference (IACC), 2010 IEEE 2nd International*, 19-20 February 2010.

[2] M. Anthimopoulos, B. Gatos, and I. Pratikakis. A two-stage scheme for text detection in video images. *Image and Vision Computing*, Volume 28, September 2010.

[3] M. S. Das, B. H. Bindhu, and A. Govardhan. Evaluation of text detection and localization methods in natural images. *International Journal of Emerging Technology and Advanced Engineering*, Volume 2, June 2012.

[4] S. Escalera, X. Baró, J. Vitrià, and P. Radeva. Text detection in urban scenes. In *Proceedings of the 12th International Conference of the Catalan Association for Artificial Intelligence*, 2009.

[5] Partha Sarathi Giri. Text information extraction and analysis from images using digital image processing techniques. *Special Issue of International Journal on Advanced Computer Theory and Engineering (IJACTE)*, Volume 2, 2013.

[6] K. Jung, K. I. Kim, and A. K. Jain. Text information extraction in images and video: a survey. *Pattern Recognition*, Volume 37, May 2004.

[7] Lukas Neumann. Scene text recognition in images and video. 31 August 2012.

[8] Y.F Pan, X. Hou, and C. L Liu. Text localization in natural scene images based on conditional random field. In *10th International Conference on Document Analysis and Recognition, 2009. ICDAR '09*, 26-29 July 2009.

[9] P.Patel and S. Tiwari. Text segmentation from images. *International Journal of Computer Applications (0975 8887)*, Volume 67(19), April 2013.

[10] T. Pratheeba, V. Kavitha, and S.R. Rajeswari. Morphology based text detection and extraction from complex video scene. *International Journal of Engineering and Technology*, Volume 2, 2010.

[11] V. V. Rampurkar, G. J.Chhajed, and S. K. Shah. Review on text string detection from natural scenes. *International Journal of Engineering and Innovative Technology (IJEIT)*, October 2012.

[12] L. SeongHun, J. H Seok, M. KyungMin, and K. JinHyung. Scene text extraction using image intensity and color information. In *Chinese Conference on Pattern Recognition CCPR 2009*, 4-6 November 2009.

[13] Y. Zhong, Z. Hongjiang, A. K. Jain, and Fellow. Automatic caption localization in compressed video. *IEEE Transactions on pattern analysis and machine intelligence*, Volume 22(4), April 2000.

[14] M. K. Y.W. K. Zhu and Q. Feihu. Using adaboost to detect and segment characters from natural scenes. In *Proceedings of First Camera-based Document Analysis and Recognition*, 2005.