International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.10 (2016), pp.407-420
http://dx.doi.org/10.14257/ijmue.2016.11.10.38

A Robust Approach for Overlay Text Localization and Extraction in Complex Video Scene

Jingfan Tang[1], Zhitao Li[2], Xingqi Wang[3], Ming Jiang[4] and Ziyang Li[5]

[1, 2, 3, 4]Institute of Software and Intelligent Technology, Hangzhou Dianzi University, Hangzhou 310018, China
[5]Hakim Information Technology Co., Ltd, Hangzhou, 310018, China
[1]tangjf@hdu.edu.cn, [2]lizhitaohappy@163.com, [3]xqwang@hdu.edu.cn,
[4]13588161992@163.com

Abstract

Overlay text in video carries important semantic clues for video information retrieval and summarization. In this paper, we propose a robust method that is able to accurately locate text lines and extract text even in complex video scene. In the text localization stage, this paper adopts the method based on corner point. First, corner detection is used to extract corners as text features from video frames. Then multi-layer filtering mechanism (MLFM) is used to locate the text lines, which consists of corners clustering, corners horizontal projection, background filtering and heuristic rules. This MLFM can effectively remove the isolated corners, locate the text lines accurately and remove the background or pseudo text lines automatically. In the text extraction stage, this paper proposed a twice binarization method that combines with polarity judgment on image. The polarity judgment was used as a guide to adjust the first binarization threshold when we perform the first binarization. After the first binarization, a main proportion of the image has been processed, and the rest will be processed by the second binarization. Experimental results show that this approach can fast and robustly locate text lines and extract text in video even under complex background.

Keywords: *Text localization; Text extraction; Polarity judgment; Corner; Twice binarization*

1. Introduction

The intelligent analysis of video data is currently in wide demand because a video is a major source of sensory data in our lives [1]. However, text is one of the most semantic information types without requiring complex computation, which carries high-level information such as program introductions, special announcements, scene locations, sports scores, date and time, news events and video contents. Text recognition in real life also has a lot of application scenarios, such as video classification, document analysis, video content based video retrieval, assistance to visually impair persons, license plate recognition, sign recognition, *etc.*

Video text recognition is mainly divided into four steps: detection, localization, extraction and recognition. Text detection is to detect whether the video frame contains text candidates. Text localization is to locate the position of the text regions in video frame accurately. The extraction is to filter out background pixels of the text candidates, so that only the text pixels are left in the text regions. Text recognition is to feed the binary image into the optical character recognition (OCR) module to produce the text. Therefore, the result of text recognition depends largely on the results of text extraction and the selection of OCR. The current commercial OCR has entered into a relatively mature stage, so text extraction is critical for high recognition rate. However, the good results of text extraction is also highly depend on accurate localization. Therefore, this paper is devoted to the study of text localization and extraction.

Text localization and extraction is actually not a newer problem in the multi-media analysis field. There are lots of methods have been proposed to locate overlay texts in video by addressing complex background or low resolution in the literature. The methods of text localization can be mainly included the methods based on color, texture and edge or gradient. Since the methods based on color [2-3], which may not be suitable for text localization due to the need of full shape of each character. The methods based on texture are developed [4,6,8] to address the problem of complex background. But these methods are low performance and computationally expensive. The methods based on edge or gradient [9-10] were developed to achieve efficiency. However, it give more false positives due that it is sensitive to the background. Therefore, most of the above methods focus on one type method rather than making good use of the advantages of a variety of methods. Existing methods of text extraction [11-13] assumed that the text color is usually light or dark. However, these assumptions can directly limit the application of text extraction based on assumption to some specific domains, *e.g.* news, captions, sports. Therefore, these methods cannot be well used for text localization and extraction in complex video scene. Overall, we proposed an effective multi-layer filtering mechanism (MLFM) for text localization and twice binarization method that combines with polarity judgment for text extraction to address complex background in video.

Compared with previously proposed methods [2-12], our work makes contribution to the following two aspects: 1) MLFM based on corners in this paper consists of corners clustering, corners horizontal projection, background filtering and heuristic rules, which enables the text to be located accurately. 2) The twice binarization method, which combines with polarity judgment of image, can do well on text extraction for the selecting of the best threshold and binarization under various conditions.

The rest of the paper is organized as follows: Section 2 is a literature of related work. It presents the detection, localization and extraction procedure in Section 3. Experimental results are presented in Section 4. It finally concludes the paper and the future work in Section 5.

2. Related Work

The proposed method belongs to the text localization and extraction, so this section provides some literature review of previous studies in text localization and extraction.

The methods of text localization are mainly classified into three types: the methods based on color, edge or texture and machine learning. 1) The method based on color assumes some constraints on text regions are satisfied such as certain sizes, uniform colors and spatial alignments. Shekar *et al.* [2] proposed a novel method based on the phase congruency model and morphology based approaches for text detection, He investigated the matching features and morphological based on a set of rules from false positive elimination and locate text by using connected component. Smitha *et al.* [3] also detect text from video frames based on skeleton matching. He obtain the skeleton by using morphology based approach and locate text by designed geometrical rules and morphological operations. It is efficient when the uniform regions are contained in video frames, it fails when there are different colors in a text lines. 2) The method based on edge or texture assumes that the backgrounds are much smoother than text regions. Therefore, it is possible to classify text regions and non-text regions according to different edge or texture intensity. However, it is still an open problem about how to reduce noises under complex backgrounds. Roy *et al.* [4] proposed a new method using fusion obtained by wavelet and gradient sub-bands to obtain text candidates. Zhang *et al.* [8] proposed video text localization method based on a fusion map of color gradients and log-Gabor filter. The method segments initially characters, and then a fusion map is constructed to fuse the results of color gradient and log-Gabor filter. Phan *et al.* [9] proposed a "gradient vector flow" model to extract the characters of text and remove false positives by the Histogram of oriented gradients feature. In [6], corners are also proposed for text detection, which are trained to produce a decision tree as the classification criteria for text characters. The above approaches may usually have a high recall rate. But it produced high false alarms due to the strong edge and texture. 3) The method based on machine learning uses features extracted from text regions and non-text regions to train neutral network or support vector machine. Samabia *et al.* [7] proposed a caption text detection method which Union of two feature vectors is used for the classification of text and non-text objects using support vector machine (SVM). Jain *et al.* [5] proposed a text detection method in natural scenes and video based on rich shape descriptors such as Histogram of Oriented Gradients, Gabor filter, corners and geometrical features of text, which train a support vector machine to classify text regions and non-text regions. Wang *et al.* [10] proposed a method combining a multi-layer neural networks with unsupervised feature learning, which can train accurately text detectors and recognizer modules. However, the performance of these methods depended on the number of training samples.

After text localization, text extraction is required to be done before text recognition. Text extraction can be divided into the methods based on threshold [11,13] and based on strokes [12]. Text can be extracted by the threshold since there is a big difference between the text and the background in the grayscale and the adjacent gradient. Otsu [13] is a text extraction method based on the threshold, which is simple in principle and have a high efficiency. However, Otsu is a method based on global thresholds, which fails to extract text exactly under complex background. Luyu *et al.* [11] proposed an adaptive threshold method, which divided the image into multiple blocks and then carried out binarization in each block. The method based on strokes is also used for text extraction. Sao *et al.* [10] proposed a strokes filter for text extraction, which enhanced the strokes and suppressed the non-text strokes by the characters of four directions formed text extraction filter.

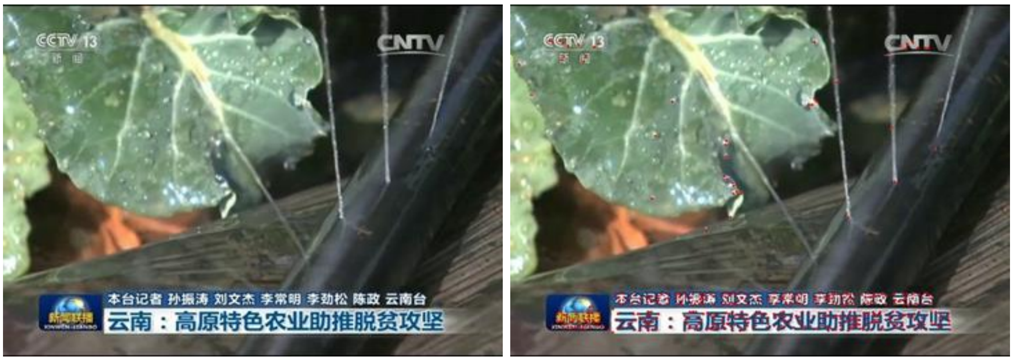
3. Our Method

The primary purpose of this paper is to solve the problem of overlay text localization and extraction under the complex background. In the stage of text detection, we choose corner as the text feature, which are not easily affected by contrast and color. In order to reduce the alarm rate of text localization, MLMF is proposed, which consists of corners clustering, corners horizontal projection, background filtering and heuristic rules. Finally, the twice binarization method combined with polarity of image can effectively enhance the effect of text extraction under complex background.

3.1. Text Detection

Text detection is the most basic step to recognize text in video frames. It is very important to choose a reliable text feature. Corners [6] are frequent and essential patterns in text regions and the advantages of corners in text regions are more orderly than the non-text regions. Therefore we use the corner point as the basic feature of the text for detection.

Corner, which usually has high curvature on the boundary of the image, is an important feature of image texture features. The text regions in the image are rich in the corners with concentrated distribution, and the arrangement of the characters is the same as that of the text. On the contrary, the corner of non-text regions is less and the distribution is stray and the random is obvious. Therefore the text regions can be detected by the density and the distribution of the corners. However, the choice of detection algorithm has a great influence on the corner detection. There are two main types: the method based on edge and the method based on grayscale. The method based on edge need to be encoded, which relies on the image extraction and the edge extraction. The two operations itself are very difficult and require complex computation, which is not good for detection. The method based on grayscale is based on the change of the neighboring gray pixels, which is effect for detection. So this paper uses the Harris corner detection algorithm based on the principle of the latter.


CCTV 13
CNTV
本台记者 孙振涛 刘文杰 李常明 李劲松 陈政 云南台
云南：高原特色农业助推脱贫攻坚

(a) (b)

Figure 1. Harris Corner Detection (a) The Original Video Frame (b) The Corner Detection Result

Harris corner detection algorithm is based on the differential operation and the autocorrelation matrix to carry out the detection of corners. Harris corner, since the differential operator can reflect the intensity of gray change in any direction, can effectively distinguish the corners and edges, so the corner detection operator has the rotation invariance. At the same time, the Harris algorithm selects the Gauss function as the detection window, which can smooth and filter image, and then detect corners of image. This method can also reduce the noises. Figure 1 shows the results of the Harris corner detection. It shows that the text regions of the corners are more dense and well-organized. But the corner of the non-text regions is relatively sparse. So the corners are a robust feature for the text localization.

3.2. Text Localization

Although text detection based on corner has many advantages, it can also produce highly false alarms for text localization especially under the complex background, which is still a challenge to locate text for correct localization rate by corner. Therefore, the MLFM is developed for localization in this paper, which consists of corners clustering, corners horizontal projection, background filtering and heuristic rules.

3.2.1. Corners Clustering

Since all of points with high curvature boundary are judged to be the corners, it contains not only lots of corners on text regions but also some corners on non-text regions. The distribution of the corners is dense on text regions but isolated on non-text regions. Therefore, we carry out corners clustering for filtering out the isolated corners in the corners distribution map, so that we can avoid the noises accumulation and improve the accuracy of the text lines localization.

Typically, it is much denser in the text regions, so we filter out the isolated corners by corner clustering to determine the distance between two corners with Euclidean distance. For any two corners, assume their positions are (x_1, y_1) and (x_2, y_2) in the image, we can calculate the distance between them.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

Where d is the distance between two corners in the image. If d is less than T (T is an empirical value), it will be regarded as the same type. It arbitrarily selects an initial corner as the first point set from corners map, and then calculates the distance between any corner within the current point set and the corner out of current set. If d is less than T, the corner will be put into the current set until there are not corners in the corners map. At this moment all corners are divided into several corners sets. Finally, corners sets whose number of corners are less than 3 will be removed.

3.2.2. Text Lines Localization

Most non-text corners have been filtered out after corners clustering. But the text extraction stage requires that the background of the text regions should be as little as possible. So we must find an effective algorithm to locate text regions accurately. We found that texts in video frames are post-production and texts is usually printed font and font is uniform. Theoretically, there is no corner between text lines. The distribution map of the corners is consistent with text position. Therefore, as a horizontal projection profile to summarize the candidate corners over rows, a projection profile method based on the number of corners is employed to separate text regions into text lines, which can directly locate the text lines in video frames. Figure 2(a) shows an example by projection based on corner. There is an obvious valley between two lines of text. We can locate the text lines accurately from video frame by choosing a proper threshold. Text line localization is illustrated as follows:

Firstly, the pixel value of the (i, j) can be expressed as $g(i, j)$.

$$g(i, j) = \begin{cases} 0 & non-corner \\ 1 & corner \end{cases} \quad (2)$$

Thus, the integral projection of the row in horizontal direction can be determined as

$$\sum_{j=1}^{L} g(i, j) \quad (3)$$

Where L is the length of row. The integral projection of each row is marked as

$$G(i) = \sum_{j=1}^{L} g(i, j) \quad (4)$$

$1 \le i \le N$, where N is the total number of rows. Average value of integral projection is expressed as

$$Average = \sum_{i=1}^{N} G(i) / line \quad (5)$$

Where line is the number of $G(i)$ that is not zero. The text lines localization algorithm is defined as:

A Horizontal Projection Based on Corners For Text Lines Localization Algorithm：

Input: Corner filtered image G

Output: Text lines

Process:

```
for $i < Height$ do
  if $upCheck()$ then
      Mark $i-th$ row as the upper border of text lines;
      if $underCheck()$ then
          Mark $i-th$ row as the lower border of text lines;
          Save the text lines images;
      end if
  end if
end for
```

Where $upCheck()$ is the function that finds the upper border of text lines. It needs to satisfy the following two conditions.

1). $(G(i) > Aver/\alpha) \bigcap (G(i+1) > Aver/\alpha) \bigcap \cdots \bigcap (G(i+n-1) > Aver/\alpha)$

2). there is at least one row in from i to $i+n-1$ which satisfies: $G(k) > Aver/\beta$ where k satisfy $i \le k \le i+n-1$. n is the threshold of line number to check upper border. $Aver$ is the height of corners histogram.

Where $underCheck()$ is the function that finds the lower border of text lines. It needs to meet the following two conditions.

1). $(G(i) < \delta) \bigcap (G(i+1) < \delta) \bigcap \cdots \bigcap (G(i+m-1) < \delta)$.

2).One row must be included from $i-th$ row to $(i+m-1)\ th$ row meeting: $G(k) < \sigma$. Where k meets $i \le k \le i+m-1$. Where m is the threshold of line number to check lower border.

Where $\alpha, \beta, \delta, \sigma$ are threshold to eliminate noises in text regions.

3.2.3. Background Filtering

The text lines can be located after the horizontal projection based on corners. Some residual background is also reserved in the two poles of text line, so it is necessary to filter out the residual background of the text lines. The traditional way is to carry out the vertical projection. But we locate text regions by corner features in this paper, which is sparser than pixels. The vertical projection based on corners are not suitable to locate the text lines. So this paper makes good use of the difference of the corner density between text regions and non-text regions to filter the background in text lines.

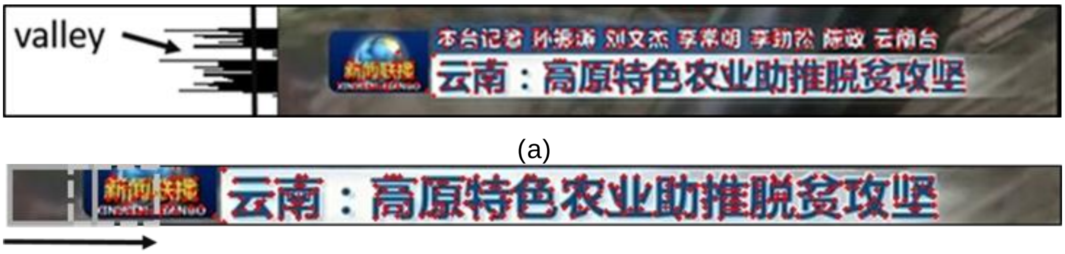
We design a size of $H \times 1.5H$ (H is the height of text line) sliding window for filtering out the background and the step-size is H from the left to the right of text line. An example has been shown in Figure 2 (b). If the number of corners in window is more than T (corner number threshold), the window will be reserved. Otherwise, the window will be discarded. Then all the reserved windows are merged by the rules that the distance between the windows is less than $2H$. And then the merged windows will be checked by heuristic rules to filter out pseudo-text lines. Finally, the merged windows refined by heuristic rules will be saved as images. Rules are defined as follows:

1. Filter out the height of text lines which is less than 8 pixels;

2. Filter out the candidate regions of minimum external matrix which is below a threshold. Text regions are usually rectangular, and the depth to width ratio is at least 1:3;

3. Filter out the 2/3 top portion of the video frame. Video content and video captions usually are in the 1/3 bottom portion of the video frames;

4. Filter out the 12 pixels lines from bottom sides of border of the video frame. The contents of this text regions are usually the program advertising, notice or other information, which is not related to the video content.



(a)

(b)

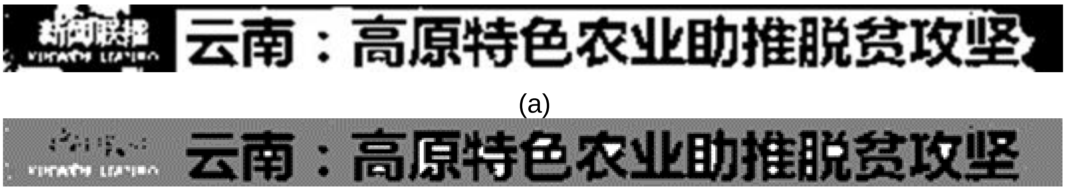
Figure 2. Text Lines Localization (a) Horizontal Projection Based on Corner Locate Text Lines (b) Remove the Background of Text Lines Based on Density of Corners

3.3. Text Extraction

Text localization is a basis step for text recognition. Text extraction is the key step to enhance the recognition rates. Text extraction means that the text of the image will be divided into the non-text and text pixels to get binary image. However, there are still many problems in text extraction under complex background at present, which mainly includes two aspects: 1) the contrast of the text lines is unpredictable such as light text and dark text. 2) The background of text is complex. A single threshold can't effectively complete the extraction. Therefore, this paper proposed a binarization method which combined with polarity judgment based on binary image.

3.3.1. Polarity Judgment Based on Binary Image

There exists a contrast color between text and non-text regions in text line. The contrast has two types. One is that the brightness of text is lighter than the brightness of background, which is named positive contrast colors. The other is that the brightness of text is darker than the brightness of background, which is named negative contrast colors. Polarity judgment can be used to judge the contrast colors. Text extraction needs to make sure the polarity of text lines which is useful to help adjust the best threshold automatically for binarization and filter out the noises of binary image. Therefore, polarity judgment is necessary for text extraction.



(a)

(b)

Figure 3. Polarity Judgment (a) Binary Image (b) The Result by Algorithm for Filling 8-adjacent Connection Based on Binary Image

In this paper, we use local Otsu method to get binary image. The pixels of four borders of binary image will be used as the seed points for eight connected domain filling algorithm. If the pixel value is the same as its neighborhood, it will be filled with 128. The result of polarity for text line image will be decided by the proportion of black pixels and white pixels. The detailed process of polarity judgment is defined as follows:

Step1. The each border of text line in image P will be extended 4 pixels and get new text lines named image $P1$;

Step2. Convert $P1$ to the gray image $P2$, use local Otsu method to get binary image $P3$.

Step3. Select the pixels of four borders of $P3$ as the initial seed pixels and put them into array $A[x, y, v]$. Where (x, y) is the position of the seed pixel in $P3$ and v is the value of pixels at (x, y) .

Step4. Select the seed points from the initial seed pixels array $A[x, y, v]$ and search the eight neighborhood from the current seed points. If the adjacent point is the same as the current initial seed point, the adjacent point will be settled 128 (128 can be changed as long as it's not 0 or 255). Then we use the current search point as the seed point to search the new seeds according to the rules of Step3 until that the point can't meet it. This means that the current initial seed point of the neighborhood search is completed.

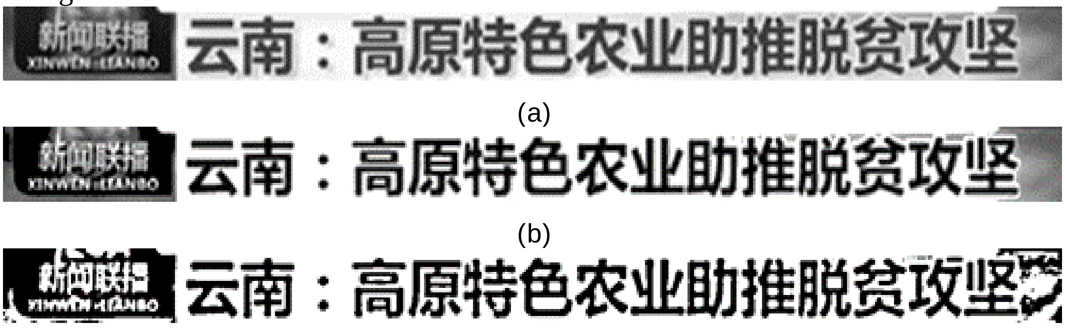
Step5. Select a new seed points from the initial seed pixels array $A[x, y, v]$. We will skip the initial seed points if its pixel value has been filled with 128 until that an initial seed point is found in $A[x, y, v]$ which hasn't been filled. Then follow the rule of Step4 to carry on the regional growth until that no seed pixels were found.

Step6. The pixel value in image only contains three types after Step5: 128, 0 and 255 in $P3$, respectively. The portions that the pixel value are 128 is the background regions. So we can judge the polarity of text line by statistic the proportion of the number of 255 ($N1$) and 0 ($N0$). The main portion are text regions and the fewer portion is non-text regions. Text is positive polarity if $N1 > N0$, otherwise Text is negative polarity if $N1 < N0$. The polarity judgment is completed.

Figure 3 (a) shows the result of the local Otsu to generate a temp binary image. Figure 3 (b) shows the result filled by 128 with algorithm for filling 8-adjacent Connection regions. Most of the background regions have been filled with 128 and only a bit of background is reserved in the closed strokes. But the proportion of backgrounds is very smaller than text regions, so it can ensure the reliability of the polarity judgment. Since this method was carried out on the binary image, itis much faster with algorithm for filling 8-adjacent Connection regions and more robust than the traditional method due to the simple binary image. The accuracy of proposed method is 97%.

3.3.2. Twice Binary

The core part of the text extraction is binarization, while the existing binarization algorithm is mainly divided into two types. One is the global threshold binarization and the other is the local threshold binarization. It is difficult to find a reasonable threshold to separate the text and non-text by the method of global threshold binarization under complex background. However it will cause too much noise through the local threshold for text extraction, which is difficult to be removed from binary image. Therefore, we adjust the threshold automatically for the first binarization through polarity judgment and then make good use of the advantages of the global threshold and the local threshold for binarization, which can solve the problem of text extraction effectively under complex background.



(a)

(b)

(c)

Figure 4. Twice Binarization (a)Text Line Gray Image (b) Local Otsu for the First Binarization (c) Niblack for the Second Binarization

First of all, we calculate the threshold by the local Otsu algorithm and then adjust the best threshold according to the results of polarity judgment for text lines image. After the first binarization is completed, Niblack's algorithm will be used for the second binarization. Figure 4 (b) shows the first binarization for original Figure 4 (a). The majority portion of image has been binarization in addition to the border of text strokes and some background of text lines are gray pixels. Figure 4 (c) shows the result of second binarization used by Niblack algorithm, whose background and the stroke are very good to be separated and strokes are clear. Twice binarization is as follows:

Step1: Convert text lines image P to the gray image G ;

Step2: Amplify 3 times for image G by bilinear interpolation and then carry out median filter;

Step3: Calculate the threshold by a size of Height x Height (Height is the height of text lines) window by Otsu from the left of image G ;

Step4: Change the threshold T according to the result of polarity judgment (The text color is black if $N0 > N1$, and then $T1 = T - N$. Conversely, the text color is white if $N0 < N1$, and then $T1 = T + N$. N is an empirical value, $N = 20$) to get the best threshold $T1$;

Step5: Carry out the first binarization. If the current gray pixel value V meets $|V - T1| > n1$ ($n1$ is an empirical value, $n1 = 30$) and $V > T1$, then $V = 255$. If $|V - T1| > n1$ and $V < T1$, then $V = 0$;

Step6: Go to Step2 if the sliding window doesn't reach the right boundary of the image G ;

Step7: Carry out the second binarization. The rest of gray pixels are completed by the algorithm of Niblack after Step6;

Step8: Image G is completed by twice binarization.

4. Experimental Results

The purpose of text recognition is to make it convenient for us to master more information in our life. The satellite televisions and CCTV brought us lots of information about life across our country. So our datasets are from these videos. They are diverse text types including different size, quality, colors, and complex background. And, all the video frames have a size of 480×360 pixels and total duration is 2 hours.

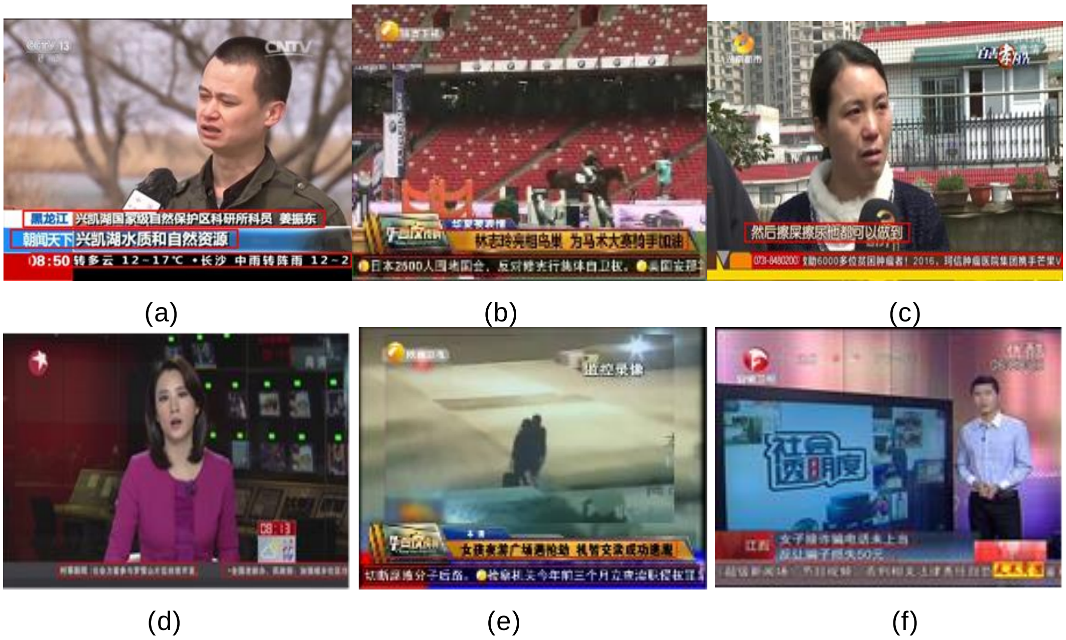
The criterion of evaluation in the text localization and extraction are different, which will be divided into two stages. The first stage focuses on the text localization evaluation, which evaluates the accuracy of the proposed algorithm in various types of video. The second phase evaluates the text extraction algorithm, which tests the validity of binarization in text extraction and the text recognition rate.

4.1. Evaluation of Text Localization

This stage is mainly to evaluate quantitatively text localization, and then compare with other methods. Experiments show that our proposed localization algorithm is robust on datasets. Figure 5 (a), (b), (c) shows some examples for locating text lines accurately in addition to the logo of TV station and other texts in the upper regions of video. Our goal is to obtain the main information of videos, which are usually in the 1/3 lower portion of the video frame. So the results of text localization are identical with our expectation. Figure 5 (d), (e), (f) shows some error examples of localization. Our algorithm failed to locate some number regions in Figure 5(d) because of the fewer corners on number regions. It also failed to locate a place in Figure 5(e), (f) due to the low resolution or be in between two text lines. Although some examples of localization failure were found in some video frames, it didn't affect the extraction for the main information of video at all.

(a) (b) (c)

(d) (e) (f)


黑龙江 兴凯湖国家级自然保护区科研所科员 姜振东
朝闻天下 兴凯湖水质和自然资源
08:50 转多云 12~17℃ •长沙 中雨转阵雨 12~2
林志玲亮相鸟巢 为马术大赛骑手加油
然后擦屎擦尿他都可以做到
监控录像
女孩夜游广场遇抢劫 执智交涉成功逃脱

(a) (b) (c)

(d) (e) (f)

Figure 5. Video Text Lines Localization Examples Figure (a), (b), (c) Text Lines Localization Positive Examples Figure (d), (e), (f) Text Lines Localization Negative Examples

Here, we take the three most widely evaluation criterion: recall rate, precision rate and speed.

1). ***Recall Rate.*** The recall rate evaluates on the proportion of text lines correctly located to total text lines in datasets. That is,

$$Recall = Lc / Tn \tag{6}$$

2).***Precision Rate.*** The precision rate evaluates on the proportion of text lines correctly located to total text lines located in datasets. Namely,

$$Precision = Lc / Ln \tag{7}$$

3). ***Speed.*** It shows how long to evaluate the text localization per frame. All the experiments are run on the same computer with 3.2Hz i5-3470 CPU.

$$Speed = T / N \tag{8}$$

Where Lc is located number of the correct text lines. Tn is the total Number of text lines in our datasets. Ln is the total number of text lines located with our method. T is the total time for text localization in total datasets. N is located number within T .

Table 1. Comparison with Others in Text Localization

Method	Precision(%)	Recall(%)	Speed(s/frame)
Ref[14]	87.47	89.56	0.22
Ref[15]	84.28	85.63	1.65
Proposed Method	89.15	92.24	0.67

To demonstrate the effectiveness of our method, we compared with algorithms proposed in [14-15]. Table 1 shows that the recall rate and accuracy of our method are higher than the other two methods. The reason is that our selection of reliable corner as text features and multi-layer filtering mechanism is more effective for locating text accurately. Even though the speed is slower than that of [14], the accuracy and recall rate of proposed method is obviously higher than that of [14].

4.1. Evaluation of Text Extraction and Recognition

This stage is to evaluate the effect of text extraction and recognition. Text extraction is directly related to the results of the text recognition, so we use the text recognition like other papers to reflect the effect of text extraction. We randomly selected 200 text lines images located by our localization method as datasets for text extraction. We will compare with the classical algorithms: Otsu [13], Niblack [16], Sauvola [17] and Wolf [18][1] algorithms. Figure 6 shows the results of text extraction. It is obvious that the Otsu's method is poor for background color variation images. Because it focuses on the global threshold by analysis the histogram-based of image. Niblack's method is better than Otsu's method, but it missed some text stokes, since it is affected by the nearby background. Since Sauvola's method make the additional assumptions for binarization, it leads to lots of thin features and holes on text strokes.The strokes of texts proposed by Wolf are relatively clearer than former methods. But the details of the text strokes are a serious problem. Figure 6 (a) shows that there is a problem of over extraction in the same image with different contrast levels. Figure 6 (b) shows an example that there is some conjoint strokes in the details for texts. But our method proposed in this paper has better results.

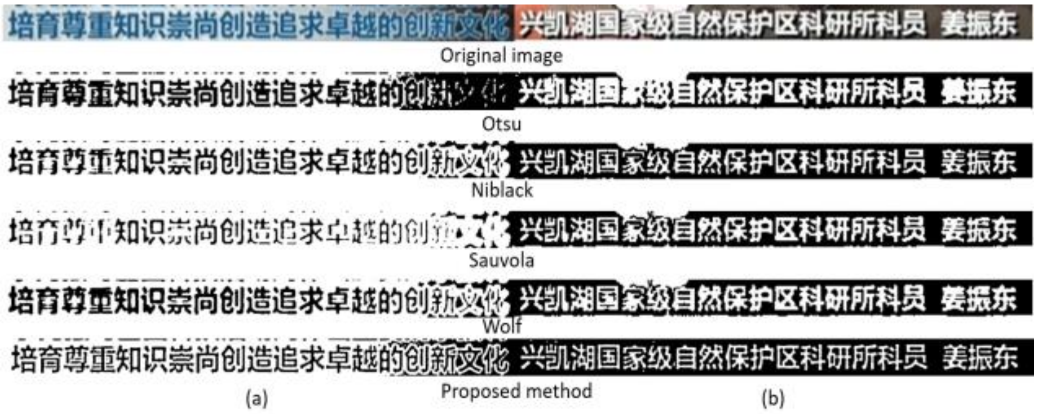

培育尊重知识崇尚创造追求卓越的创新文化
兴凯湖国家级自然保护区科研所科员 姜振东
Original image
Otsu
Niblack
Sauvola
Wolf
Proposed method
(a)
(b)

Figure 6. Comparison of Text Extraction Method

We evaluate the effect of text extraction by recognition rate. We will feed the text extracted by the above five methods into the OCR module (this paper uses the open source Tesseract-OCR[2]) for text recognition. We use the word recognition rate as the evaluation criteria for the evaluation of recognition efficiency, where recognition rate is the number of correct word recognized by OCR accounts for the percent of the total number of words.

Table 2. Word Recognition Rate in Different Extraction Methods

Method	Otsu	Niblack	Sauvola	Wolf	Our Method
Recognition(%)	55	52	60	67	72

Table 2 shows the results of word recognition rate in different extraction methods. Our recognition rate is far better than others, which benefits from the polarity judgment

[1] http://liris.cnrs.fr/christian.wolf/software/binarize/index.html
[2] http://code.google.com/p/tesseract-ocr/

combined with the twice binarization for extraction. Otsu is too single in thresholds, and Niblack algorithm generate too much noises. Sauvola and Wolf algorithm have a better effect, but the strokes are lost due to over extraction.

5. Conclusions and Future Work

In this paper, an effect text location and extraction algorithm are proposed for complex background video. The purpose of this paper is to improve overlay text recognition rate under complex background in video. For text localization, we proposed a MLFM which can accurately locate the overlay text lines and can effectively overcome the low accuracy rate for the traditional text localization based on corners. For text extraction, the twice binarization method combine with polarity judgment based on binary image was proposed to solve the problem under complex background in video. The twice binarization is composed of global and local binarization, which can retain the overall and detail of the text strokes. Polarity judgment based on binary images, which can judge automatically the text color, can effectively improve twice binarization on best threshold choice.

Although we have completed the work of text localization and extraction under complex background images in video, the proposed algorithm in this paper is only for the overlay text rather than the scene text in video. Scene texts also have great importance in our life such as the road name, product description, scenic spots introduction, *etc.* Therefore, our future work will focus on the scene text recognition.

Acknowledgments

This work is supported by national Science and Technology Support Program of China (No. 2014BAH23F01-2) and Zhejiang Provincial Technical Plan Project (No. 2014C01051, 2015C33003, 2013C01113)

References

[1] X. C. Yin, Z. Y. Zuo, S. Tian and C. L. Liu, “Text Detection, Tracking and Recognition in Video: A Comprehensive Survey”, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, vol. 25, no. 6 **(2016)**, pp. 2752-73.

[2] B. H. Shekar and M. L. Smitha, “Phase congruency and morphology based approach for text localization in videos”, International Conference on Advances in Computing, Communications and Informatics, **(2015)**.

[3] B. H. Shekar and M. L. Smitha, “Skeleton Matching based approach for Text Localization in Scene Images”, Computer Science, **(2015)**.

[4] S. Roy, P. Shivakumara, P. P. Roy and C. L. Tan, “Wavelet-Gradient-Fusion for Video Text Binarization”, Pattern Recognition (ICPR), 2012 21st International Conference on, **(2012)**, pp. 3300-3.

[5] Jain, X. Peng, X. Zhuang, P. Natarajan and H. Cao, “Text detection and recognition in natural scenes and consumer videos”, ICASSP 2014 - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, **(2014)**, pp. 1245-9.

[6] X. Zhao, K. H. Lin, Y. Fu, Y. Hu, Y. Liu and T. S. Huang, “Text From Corners: A Novel Approach to Detect Text and Caption in Videos”, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, vol. 20, no. 3, **(2011)**, pp. 790-9.

[7] S. Tehsin, A. Masood, S. Kausar and Y. Javed, “A Caption Text Detection Method from Images/Videos for Efficient Indexing and Retrieval of Multimedia Data”, International Journal of Pattern Recognition & Artificial Intelligence, vol. 29, no. 1, **(2013)**.

[8] Z. Zhang, W. Wang and K. Lu, “Video Text Extraction Using the Fusion of Color Gradient and Log-Gabor Filter”, International Conference on Pattern Recognition, **(2014)**, pp. 2938-43.

[9] T. Q. Phan, P. Shivakumara and C. L. Tan, “Text detection in natural scenes using Gradient Vector Flow-Guided symmetry”, Marine Chemistry, vol. 35, no. 1-4 **(2012)**, pp. 3296-9.

[10] T. Wang, D. J. Wu, A. Coates and A. Y. Ng, “End-to-end text recognition with convolutional neural networks”, International Conference on Pattern Recognition, **(2012)**, pp. 3304-8.

[11] M. R. Lyu, J. Song and M. Cai, “A comprehensive method for multilingual video text detection, localization, and extraction”, IEEE Transactions on Circuits & Systems for Video Technology, vol. 15, no. 2, **(2005)**, pp. 243-55.

[12] T. Sato, T. Kanade, E. K. Hughes and M. A. Smith, "Video OCR for Digital News Archive", IEEE International Workshop on Content-Based Access of Image and Video Database, 1998. Proceedings, **(1998)**, pp. 52-60.

[13] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms", Systems Man & Cybernetics IEEE Transactions on, vol. 9, no. 1, **(1979)**, pp. 62-6.

[14] A. Saracoglu and A. A. Alatan, "Automatic Video Text Localization and Recognition", Journal of Supply Chain Management, vol. 44, no. 1, **(2006)**, pp. 1-4.

[15] M. Anthimopoulos, B. Gatos and I. Pratikakis, "A two-stage scheme for text detection in video images", Image & Vision Computing, vol. 28, no. 9, **(2010)**, pp. 1413-26.

[16] W Niblack, "An introduction to digital image processing", Strandberg Publishing Company, **(1985)**, pp. 72.

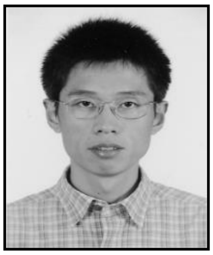
[17] J. Sauvola, T. Seppanen, S. Haapakoski and M. Pietikainen, "Adaptive document binarization", International Conference on Document Analysis and Recognition, **(1997)**, pp. 147-52.

[18] W. J. Jolion and F. Chassaing, "Text localization, enhancement and binarization in multimedia documents", vol. 2, **(2002)**, pp. 21037.

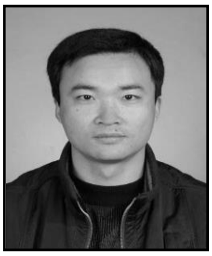
Authors

Jingfan Tang, is an associate professor in Hangzhou Dianzi University. He received his PhD in computer science from Zhejiang University in 2005. His research interests include software quality assurance, project management, network information security, *etc.*

Zhitao Li, currently is a postgraduate of College of Computer Science at Hangzhou Dianzi University. His research area is image processing.



Xingqi Wang, received his Bachelor and Master degree from Harbin Institute of Technology in 1997 and 1999, respectively, and Ph. D degree from Zhejiang University in 2002. He is an associate professor in School of Computer Science, Hangzhou Dianzi University, China. As a researcher, he visited CERCIA, University of Birmingham, UK from 2005 to 2006. His research interests include intelligent software analysis and testing, multimedia content analysis.



Ming Jiang, received the Ph.D. degree in computer science from Zhejiang University in 2004, and currently is a professor of College of Computer Science at Hangzhou Dianzi University. His research areas include network virtualization, software defined network and Internet QoS provisioning.

Ziyang Li, received bachelor's degree from Zhejiang University of Technology in 2008, and currently is the technical director for Hakim Information Technology Co., Ltd, mainly engaged in cloud data center product research and development.