Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 2187647, 11 pages
http://dx.doi.org/10.1155/2016/2187647

Research Article

A Method of Effective Text Extraction for Complex Video Scene

Zhe Guo,[1] Yuan Li,[1] Yi Wang,[1] Shu Liu,[1] Tao Lei,[2,3] and Yangyu Fan[1]

[1]*School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China*
[2]*School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China*
[3]*College of Electrical & Information Engineering, Shaanxi University of Science & Technology, Xi'an 710021, China*

Correspondence should be addressed to Zhe Guo; guozhe@nwpu.edu.cn

Received 26 January 2016; Accepted 12 June 2016

Academic Editor: Erik Cuevas

Text information contains important information for video analysis, indexing, and retrieval. Effective and efficient text extraction has been a challenging topic in recent years. Focusing on this issue, a text extraction method for complex video scene is proposed in this paper. Multiframe corner matching and heuristic rules are combined together to detect the text region candidates, which solves the issue of Harris corner filtration for complex video scene and also improves the detection accuracy using multiframe fusion. Local texture description is then used for similarity evaluation judged by SVM. Experimental results for 4 different types of 395-frame video images show the effectiveness of the proposed method compared with 5 existing text extraction methods.

1. Introduction

In recent years, image- and video-based multimedia information has been playing an increasingly important role in the fields of information exchange and services. The content-based retrieval is an important method to manage and search for the massive multimedia information [1]. In the field of content-based multimedia retrieval, the correct identification of text from images and video will lay a strong foundation for achieving the proper retrieval result. Therefore, how to extract text from the complex background becomes a crucial step for understanding and retrieving images and videos.

Generally, there are two main parts for text extraction: the text region detection and the text segmentation. The existing methods of text region detection can be divided into four categories [2]: edge-based detection, texture-based method, connected region-based method, and machine learning method. Method based on edge detection [3] detects the edges of the image by the edge detection operator, then filters the edges or aggregates the candidate text regions, and finally filters out text regions by defining some heuristic rules. This method, although it has quite high efficiency, possesses weak robustness under the disturbance of complex background. Texture-based method [4] judges whether pixel points or pixel blocks belong to the text by using the image texture features. Such method can effectively detect character region under the complex background but has low operational efficiency since it has to deal with the differential operation of the whole image. Method based on connected region [5] using image segmentation or color clustering methods extracts the same color text from the background. The premise of this approach is that the characters share the same color. However, when the complex background regions in the image have similar color as the text, the test results are not satisfactory. Machine learning based method [6] classifies text blocks and nontext blocks by constructing the mechanism of learning. Since such methods need to select samples in order to train the learning machine for classification [7], the similarity between training sample sets and test sample sets is not high enough to perform ideal detection results.

Because the detected text region contains a complex background, the text needs to be split out of it for further applications. The prevailing text segmentation methods mainly use the text color and partial space information, which can be roughly classified into the following three categories: threshold method, unsupervised clustering method, and the method based on a statistical model [8, 9], whereas the

above methods apply only to the grayscale text blocks with simple background. When it comes to the background that contains the same or similar color component as the text, there would exist misclassification, or when the number of kernel functions of statistical model is difficult to determine, the text region extraction from complex video scene does not perform efficiently.

Text superimposed in complex background video images is superimposed directly on top of the image. Thus, detected text blocks typically contain some unpredictable complex image backgrounds, which will cause obstacles for the segmentation. The diversity of the color and texture of complex background makes it difficult to estimate the text color. Meanwhile, the background of segmented text blocks only contains a small fragment of the whole original background image, and therefore information contained is limited due to the fragmentary texture, which cannot be easily described by model construction. Through the above analysis, this paper proposes a text extraction method for complex video scene. The proposed method includes two primary steps, which are text region coarse detection based on multiframe corner matching and heuristic rules and the video text region extraction under complex background based on texture and SVM aiming to position the text region precisely. Specifically, multiframe corner matching is mainly used to solve the issues of Harris corner filtration for complex background video scene. The method is based on the relationship between consecutive multiframe images, with the aid of the temporal redundancy of the video text, using multiframe fusion to improve the accuracy of text detection. Using heuristic rules to filter the candidate text regions, heuristic rules can change according to the type of scene, which can enhance the efficiency of the algorithm and also reduce the false alarm rate in some extent. LBP histogram is used to describe the local texture of the image, the similarity tolerance between images is then judged based on SVM, and finally text in complex video scene is extracted accurately.

In the experimental section, four different types of 395-frame video images including movie, news, sport, and cartoon are used for experiments, and the existing five methods are compared with the proposed one under three evaluation criteria: the text extraction accuracy, false alarm rate, and the recall rate. Experimental results show the effectiveness of the proposed method. Based on the detailed algorithm analysis, the effect of the whole system accuracy and improvement suggestions are then provided.

2. Text Region Coarse Detection Based on Multiframe Points Matching and Heuristics Rules

2.1. The Harris Corner Detection. Corner [10] is an important feature of image texture, usually defined as the qualified high curvature points on the boundary of the image. Corner can be found at the edges and contours, which is the location of violent variation of image brightness or location of curve maxima of curvature at image edge and also is independent from text features as the font color and font size.

Harris operator is a point feature extraction operator based on gray scale proposed by C. Harris and M. J. Stephens on the basis of Moravec algorithm [11]. This operator is inspired by the self-related function in signal processing and giving matrix M about the self-related coefficient, which has feature values as the first-order curvature of self-related function. If both the curvature values are high, the point is considered to be the corner point. The principle of corner detection in images can be described as if the offset in either direction at a point in the image will cause significant grayscale changes; then the point is a corner point.

The specific steps of Harris corner detection algorithm for video images are shown as follows:

(1) Convert the video image to grayscale:

$$I(x, y) = 0.299R + 0.587G + 0.114B. \quad (1)$$

(2) Calculate the correlation matrix M:

$$M = G(\bar{s}) \otimes \begin{bmatrix} g_x & g_x g_y \\ g_x g_y & g_y \end{bmatrix}, \quad (2)$$

where g_x is the gradient of x, g_y is the gradient of y, and $G(\bar{s})$ is the Gauss template.

(3) Calculate the Harris corner response of each pixel:

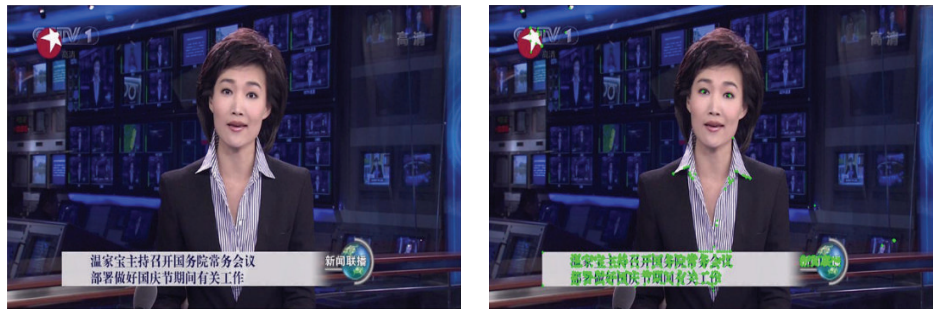
$$R = \det(M) - k \cdot \mathrm{tr}^2(M), \quad k = 0.04, \quad (3)$$

where det is the determinant of the matrix, tr is the straight stitch of the matrix, and k is the default constant.

(4) Within the scope of the Gaussian window to find the maximum point, if the Harris corner response is greater than the threshold, the maximum point would be considered as the corner point.

Since Harris operator only involves the first-order difference and filtering of image grayscale, the calculation is quite simply without threshold comparison; therefore, the whole process is highly automatic. One frame of the news video scene is processed by Harris corner detection, and the result is shown in Figure 1.

2.2. Corner Filtering Based on Multiframe Corner Matching. Since all the border points owing high curvatures in the image would be judged as corners, there might be some corners belonging to background, outside the text corners in the corner distribution image. In order to solve this problem, isolated corner filtering method is generally used to reduce noises in text extraction procedure [12]. For video images with complex background, simply using corner density filtering method is difficult to overcome the interference of complex background. To solve the problem of complex background video image filtering, this paper proposed an effective method based on the relationship between consecutive frame images and the temporal redundancies of the video text, utilizing multiframe integration strategy to improve the accuracy of text detection.



(a) The original news video image (b) Harris corner detection result

Figure 1: Results of Harris corner detection.

The similar subtitles' text in TV video will appear adjacently in several continuous video frames. And the relative location of subtitle text is fixed, so the pixels within text region will not have big changes. However, the movement of background leads to big changes in background pixel, which can be used for background filtering. In this paper, we propose a filtering method of video image corners based on multiframe corners matching under complex background. The specific procedure of the algorithm is described in the following steps.

Step 1. Detect corners of two adjacent video frames based on Harris operators, respectively; corner set of each frame consists of zero or several background corners and characters corners. The corner sets of two adjacent frames f_i and f_{i+1} can be presented as

$$\begin{aligned} C_i &= \{\text{corners\_background} \subseteq f_i, \text{corners\_character} \\ &\quad \subseteq f_i\} \\ C_{i+1} &= \{\text{corners\_background} \subseteq f_{i+1}, \text{corners\_character} \\ &\quad \subseteq f_{i+1}\}. \end{aligned} \tag{4}$$

Step 2. Find out the corner set C_{com} including both f_i and f_{i+1} presented as

$$\begin{aligned} C_{\text{com}} = C_i \cap C_{i+1} = \{&\text{corners\_background} \subseteq f_i \\ &\cap f_{i+1}, \text{corners\_character} \subseteq f_i \cap f_{i+1}\}. \end{aligned} \tag{5}$$

Step 3. For corner set C_{com}, define a sliding window w of size 15×15, and make the window slide in directions of x and y, respectively. The sliding step length is 5. Scan the entire corner distribution map, and then calculate the corner density inside the window. If the corner density is below a certain threshold, the corner of sliding window center will be removed. The set of filtered corner made as C_{filter} can be presented as

$$C_{\text{filter}} = \{c_i \mid \text{Den}(w(c_i)) \geq \text{Den}_{\text{threshold}},\ c_i \in C_{\text{com}}\}, \tag{6}$$

where c_i is a certain corner, $\text{Den}(w(c_i))$ is the corner density of window center c_i, and $\text{Den}_{\text{threshold}}$ is the threshold.

The corner filtering algorithm based on multiframe matching processes each pixel accurately in matching progress. Therefore, the location of objects in two adjacent frames of a video image will be detected, even if it moved one pixel unit distance. Effectively, our method has the ability to solve interference problem that the text region detection is vulnerable to complex background using only one single image.

2.3. Text Regions Detection under the Guidance of Heuristics Rules. Corner filtering method based on multiframe corner matching has a good performance on filtering the corners generated by the video background, involving a relative static text with a moving background. For the video subtitles with linear motions, we need to define a modified algorithm and heuristic rules [13] further improving the accuracy of text region detection. Modified algorithm is described as follows.

(1) According to the motion direction of text, sort the corner set C_{com} of both video images according to the coordinate values of x (for the text horizontal motion) or the coordinate values of y (for the text vertical motion) presented as

$$\begin{aligned} C_{\text{com}} &= \text{sort}(c_i, \text{Coor}(c_i, x)); \\ &\qquad \text{for horizontal movement} \\ C_{\text{com}} &= \text{sort}(c_i, \text{Coor}(c_i, y)); \\ &\qquad \text{for vertical movement,} \end{aligned} \tag{7}$$

where $\text{sort}(c_i,)$ means sorting corners c_i in set C_{com}, $\text{Coor}(c_i, x)$ is the coordinate value x of c_i, $\text{sort}(c_i, \text{Coor}(c_i, x))$ means sorting corners c_i according to the coordinate value of x, and $\text{sort}(c_i, \text{Coor}(c_i, y))$ means sorting corners c_i according to the coordinate value of y.

(2) For the text with horizontal motion, find out the corners with same coordinate value of x from the point sets of two adjacent video frames; calculate the difference of coordinate values of y. If the number of corners with same coordinate value of x and fixed coordinate difference of y is greater than a certain threshold and moreover the qualified corners distribute concentratively in the same region, then the point sets of two adjacent video frames can be made

For horizontal movement
  Find $C_{\text{mid}} = \{c_i,\ i = 1, 2, \ldots, m\}$, that is, $\text{Coor}(c_i, x)$ is same, $c_i \in C_{\text{com}}$;
  Compute Difference($\text{Coor}(c_i, x)$, $\text{Coor}(c_{i+1}, x)$);
  If (Difference($\text{Coor}(c_i, x)$, $\text{Coor}(c_{i+1}, x)$) ≥ Threshold) && (c_i, c_{i+1} in same unit region)
    Keep c_i, c_{i+1} in C_{mid};
  else
    Delete c_i, c_{i+1} from C_{mid}.
END

Algorithm 1

(a) Corners after filtering (b) Background-free corner map

Figure 2: Filtering results of heuristic rules.

as text candidate regions. For the text with vertical motion, the method is similar. The pseudocode of the above algorithm is described as in Algorithm 1.

After the modified method, we also define heuristic rules [13] to carry out further filtering of the candidate text regions. Heuristic rules are described as follows.

(1) Each minimum external matrix of candidate text regions is calculated. Count heights of all minimum external matrix. Make the height which appears most frequently as a standard. Candidate text regions with greater height than the standard will be filtered.

(2) Eliminate the candidate text regions with the ratio of height and width of the minimum external matrix below a certain threshold.

Through the multiframe corners matching and heuristic rules filtering, the results of Figure 1 are shown in Figure 2(a). The corresponding background-free corner map is shown in Figure 2(b).

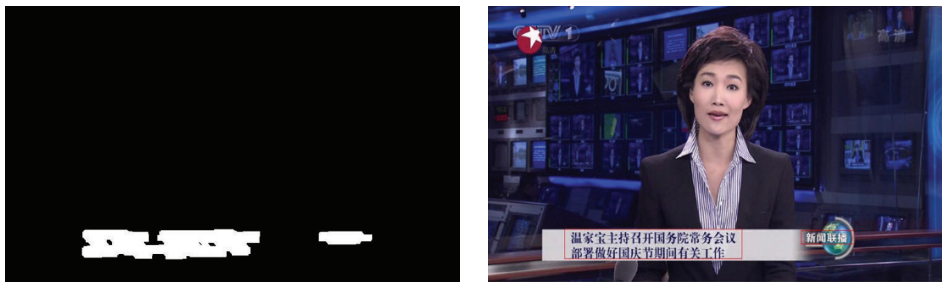
With the morphological operation to corners with no background, we can get the approximate outline of the text region. External rectangle of the outline is then found out and the text region can be finally extracted on the original image. Results are shown in Figure 3.

Heuristic rules have the advantage of improving the algorithm efficiency and also reducing the rate of false alarms, as the rules can change according to the style of the video images.

3. Text Extraction Based on Texture and SVM

LBP [14] (Local Binary Pattern) is an operator used to describe local image texture features. Therefore, LBP histogram can be used to describe the local image texture, further determine the similarity between images based on similarity measurement function, and finally complete precise text region extraction. Based on the above analysis, this paper adopts the model of uniform LBP to calculate LBP histogram of the text region extraction results obtained from the previous section. In this way, feature dimension can be reduced from the original 256 dimensions to 59 dimensions, which greatly reduce the data complexity, while keeping the effective characteristics of the data.

For the calculated LBP histogram, LIBSVM [15] is used for similarity measurement of text region and finally gives the text extraction results. We choose the RBF (Radial Basis Function) as the kernel function of SVM. Before the SVM training, we need to select parameters of SVM model, which is, namely, how to determine the optimal parameters set $\{C, Y\}$ of the Gaussian radial kernel function. Among them, C presents error penalty factor, and Y presents the Gaussian radial parameter. In our experiment, the training set is divided into 5 groups. By crossing validation of the training set among the 5 groups and using grid-search method for finding the optimal parameter set in $C = \{2^0, 2^1, \ldots, 2^{10}\}$ and $Y = \{2^{-11}, 2^{-5}, \ldots, 2^{-1}\}$, we can then obtain the value 128 of C and the value 0.015625 of Y. Finally, according to the parameters above, we generate the SVM classifier model by training the entire training set.



(a) The morphological operation results (b) Extraction results of text region

Figure 3: Extraction results of text region.

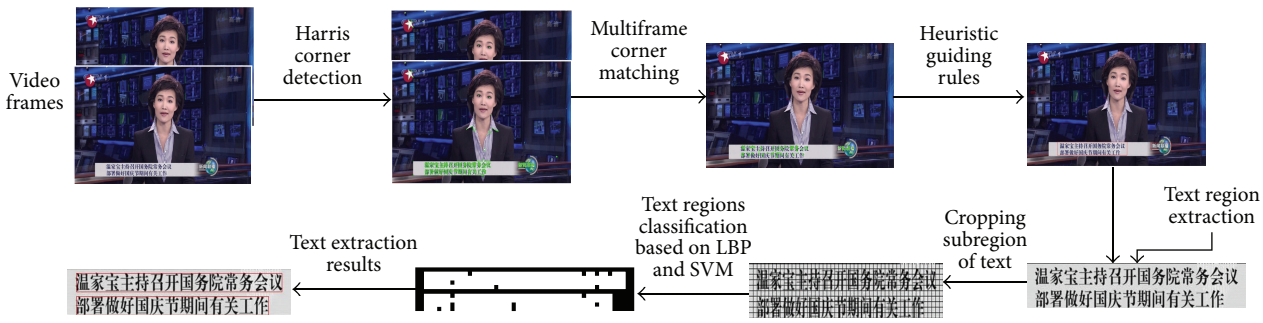
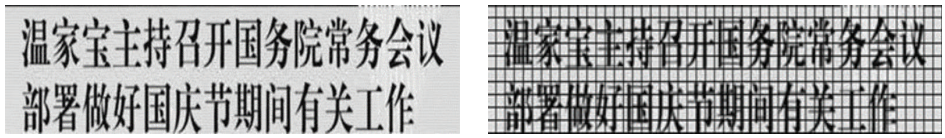


Figure 4: The flowchart of proposed algorithm.



(a) The result of cropped text region (b) Subregions of cropped text

Figure 5: Cropping results of text region.

4. Experimental Results and Analysis

4.1. Algorithm Description. In this paper, the flowchart of the proposed text extraction method for complex video scene is shown in Figure 4. Our method consists of two parts: one is the text region coarse extraction based on multiframe corner matching and heuristic rules and the other is the text extraction based on texture and SVM. Finally, the subtitle text in the video can be accurately extracted.

In terms of implementation, firstly, according to the text region labeling results of the second section, we crop the text region in the original image, convert the color image into grayscale image, and then adjust the resolution of the cropped text region to 500×100 and divide it into 50×10 subregions. Finally, we compute the LBP histogram of each region and obtain the LBP histogram of 59 dimensions as test data. For the SVM training samples, we download video images that contain Chinese subtitles from the Internet and get training test regions manually. The sample data is obtained using the above preprocessing. Class labels such as text region with "1" and the nontext region with "0" are marked manually. With all the above preparations, we start the training and get results. For the SVM training step, the quantity and quality of training samples directly determine the result of classification. Theoretically, within certain limitations, the more the training samples are, the better accuracy and robustness the classification would have. Consequently, we select 400 video frames that contain different Chinese subtitles including four different kinds such as movie, news, sport, and cartoon to complete the training.

4.2. Text Region Extraction Results. According to the results in Figure 3(b), the text region is cropped to convert into grayscale image; the result is shown in Figure 5(a). The resolution of the cropped text region is then adjusted to 500×100 and divided into 50×10 subregions, as shown in Figure 5(b).

(a) SVM classification result (b) Text region extraction results

Figure 6: Accurate text extraction results.

Table 1: Text extraction accuracy of four different kinds of video images.

Video type	Number of text regions	Accuracy (%)					
		Otsu [17]	Sato et al. [18]	Lyu et al. [19]	Song et al. [20]	Shivakumara et al. [21]	Our algorithm
Movie (121 frames)	432	60.9	69.8	86.9	72.4	78.6	92.1
News (85 frames)	261	62.1	70.3	89.3	75.3	79.4	92.4
Sport (103 frames)	368	41.0	59.4	84.1	57.2	64.1	89.8
Cartoon (86 frames)	254	68.7	81.2	90.7	81.8	88.2	96.5
Total	1315	58.2	70.1	87.8	71.7	77.6	92.7

According to the training results to classify the test data shown in Figure 5, the classification results are shown in Figure 6(a); the white part is represented in the text region. The final text extraction results are shown in Figure 6(b) as drawing an external rectangle on the white part.

Results for other video images including four different types of movie, news, sport, and cartoon by our proposed method are presented in Figure 7 (left is the coarse result, and the right gives accurate text extraction results).

As shown in Figure 7(a), in the final text region extraction results of the news video, neither the character "%" in the text "1%" at the bottom of the screen is marked, nor is the time text region. The program logo in the lower right corner of the movie video in Figure 7(b) also fails to be marked, because the fonts and text layout are quite different from traditional subtitles. In Figure 7(c), although the sport game video contains TV station logo, program title, subtitles, and a variety of complex texts, our method can still successfully extract all the text regions. In Figure 7(d), the captions and background of the cartoon video are noticeably distinguished from each other; therefore, the text regions are also able to be correctly extracted by our approach.

4.3. Comparison of the Algorithm Efficiency. In this section, we evaluate the algorithm performance of the text region extraction in the video image. Particularly, the text region here refers to the subtitle text added in the video production period. There are four evaluation standards for video text extraction, which are text extraction accuracy, false alarm rate, recall rate, and character recognition accuracy [16]. In this paper, our algorithm is aiming to extract the text region in video images. Therefore, we use the first three evaluation standards for performance evaluation. For comparison, we compared our method with five existing methods: Otsu's method [17], Sato's method [18], Lyu's method based on edge detection [19], Song's method [20], and Shivakumara's method [21].

As the unit of text blocks, the text extraction accuracy, false alarm rate, and recall rate are defined as

$$\begin{aligned}
&\text{Text Extraction Accuracy} \\
&\quad = \frac{\text{Numbers of CTR}}{\text{Numbers of CTR} + \text{Numbers of ICTR}} \times 100\% \\
&\text{False Alarm Rate} \\
&\quad = \frac{\text{Numbers of FTR}}{\text{Numbers of TRVS} + \text{Numbers of FTR}} \times 100\% \\
&\text{Recall Rate} = \frac{\text{Numbers of CTR}}{\text{Numbers of TRVS}} \times 100\%,
\end{aligned} \tag{8}$$

where CTR means correct extracting text regions, ICTR means incorrect extracting text regions, FTR means false extracting text regions, and TRVS means text regions existing in video scene.

Experimental results selected of four types of video data are shown in Tables 1, 2, and 3. It can be found that the text extraction method for complex video scene presented in this paper possesses an average accuracy rate of 92.7%, the average recall rate of 82.8%, and average false alarm rate of less than 6.1%, and all of these three indicators are higher than the above-mentioned five methods. In contrast, Lyu's method [19] is comparatively better than the other four methods; however, when it comes to do the extraction of the text data used in our experiments, it presents an average accuracy rate of 87.8%, recall rate of 80.2%, and false alarm rate probability of 6.7%, which are all below the level of our method results. The statistics of experimental results demonstrate that the precision and recall rate of the proposed algorithm are much higher than the existing five methods.

(a) News video image

(b) Movie video image

(c) Sport video image

(d) Cartoon video image

Figure 7: Text extraction results for four different types of video images.

Table 2: Text extraction false alarm rate of four different kinds of video images.

Video type	Number of text regions	False alarm rate (%)					
		Otsu [17]	Sato et al. [18]	Lyu et al. [19]	Song et al. [20]	Shivakumara et al. [21]	Our algorithm
Movie (121 frames)	432	9.8	9.0	6.9	8.1	7.5	6.2
News (85 frames)	261	8.9	7.9	6.3	7.3	6.8	5.8
Sport (103 frames)	368	12.8	10.6	9.2	9.9	9.6	8.7
Cartoon (86 frames)	254	7.0	6.1	4.3	5.5	4.8	3.6
Total	1315	9.6	8.4	6.7	7.7	7.2	6.1

Table 3: Text extraction recall rate of four different kinds of video images.

Video type	Number of text regions	Recall rate (%)					
		Otsu [17]	Sato et al. [18]	Lyu et al. [19]	Song et al. [20]	Shivakumara et al. [21]	Our algorithm
Movie (121 frames)	432	63.8	67.5	85.3	70.7	81.4	87.6
News (85 frames)	261	57.5	60.2	78.4	65.6	73.9	81.8
Sport (103 frames)	368	45.9	49.8	67.3	57.5	62.4	69.4
Cartoon (86 frames)	254	68.6	73.2	89.6	78.5	85.6	92.5
Total	1315	58.9	62.7	80.2	68.1	75.8	82.8

From the results shown in Tables 1, 2, and 3, it can be found that the typesetting of title and subtitles in news video is well standardized, with less text species and rarely art effects processing, while text and background bring out a strong difference and also last longer in video, so the extraction of text in news is relatively easy; accurate rate and recall rate of text detection are both high. Since the texts in movie and cartoon video are mainly located in the lower screen, with nice text specification, and the color of cartoon background is much simpler compared to movie, the text extraction accuracy rate for cartoon is higher than movie. Generally, sport video is much more complex which often contains banner texts which appeared randomly, complex background, and more text types and effects, so text extraction accuracy and recall rate are lowest in the 4 different videos.

4.4. Algorithm Detail Analysis. The proposed text extraction method for complex video scene consists mainly of four parts, corner coarse detection based on Harris (denoted as corner detection (CD)), corner filtering based on multiframe corner matching (denoted as corner filtering (CF)), text regions coarse detection under heuristics rules (denoted as text regions coarse detection (TRCD)), and text regions fine detection based on texture and SVM (denoted as text regions fine detection (TRFD)). The performance of each step will affect the final accuracy of the text extraction. Therefore, this part uses the statistical method to analyze detail performance of the proposed algorithm.

The selected four types of video data (totally 395 frames) are used for detailed experiments. First of all, we manually obtain all the text regions of 395 frames as the benchmark; the results of each step are then compared with the reference. For corner detection and corner filtering, the detection accuracy is calculated by the percentage of corner points which fall in the benchmark text regions. For two text regions detection steps, the detection accuracy is calculated by the coincidence rate of detected text regions and the benchmark. Experimental results are shown in Figure 8.

From the results, it can be seen that the CD step is based on Harris, as part of the detected corners for this step outside the text regions; therefore the detection accuracy is not high, especially for the sport video data which contain very complex background, and the average accuracy rate for 103 frames is only about 45%. Consequently, the corner filtering (CF) step is necessary. Through the comparison of experimental results of four different types of video in Figures 8(a), 8(c), 8(e), and 8(g), it can be found that, by the proposed corner filtering step based on multiframe corner matching, corner detection accuracy rate increased by 27%, which ensures the good input of the text regions detection step. And for comparison of the next two steps, TRCD and TRFD, TRCD based on heuristic rules, the average detection rate for movie, news, sport, and cartoon video data is 83.5%, 76.2%, 63.7%, and 87.4%, respectively. TRFD step is used for text regions fine detection; the average detection accuracy rate is 5% higher than TRCD. By comparing the results of TRCD and TRFD in Figures 8(b), 8(d), 8(f), and 8(h), it can be found that the TRCD step has a greater effect on the total system accuracy than the TRFD step. For the theoretical analysis, TRFD step used the results of TRCD step and extracted texture feature for SVM learning. Definitely, the learning results will affect the system accuracy; however, the input of the SVM, which means the results of TRCD step, has more influence on the whole learning system.

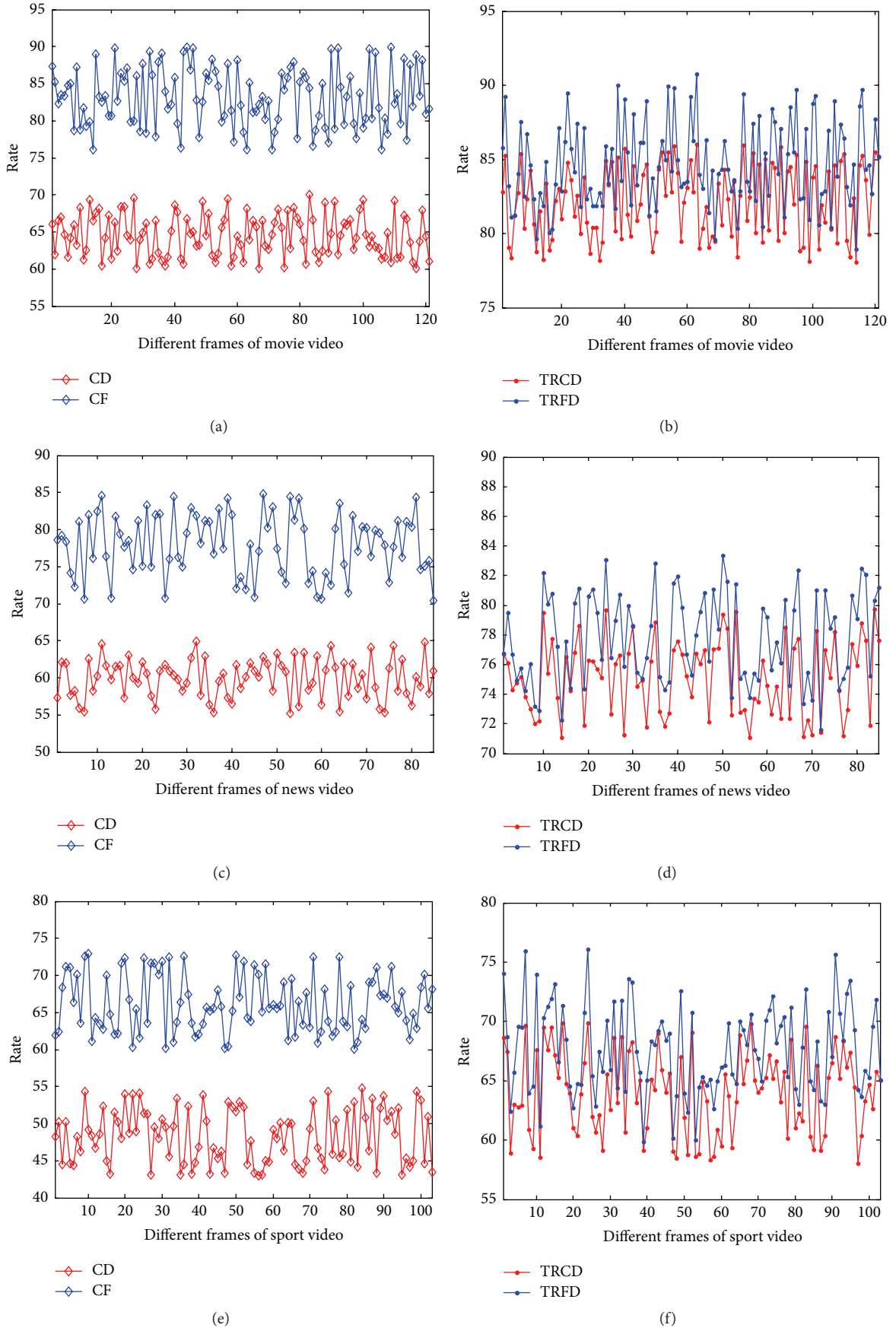


Figure 8: Continued.

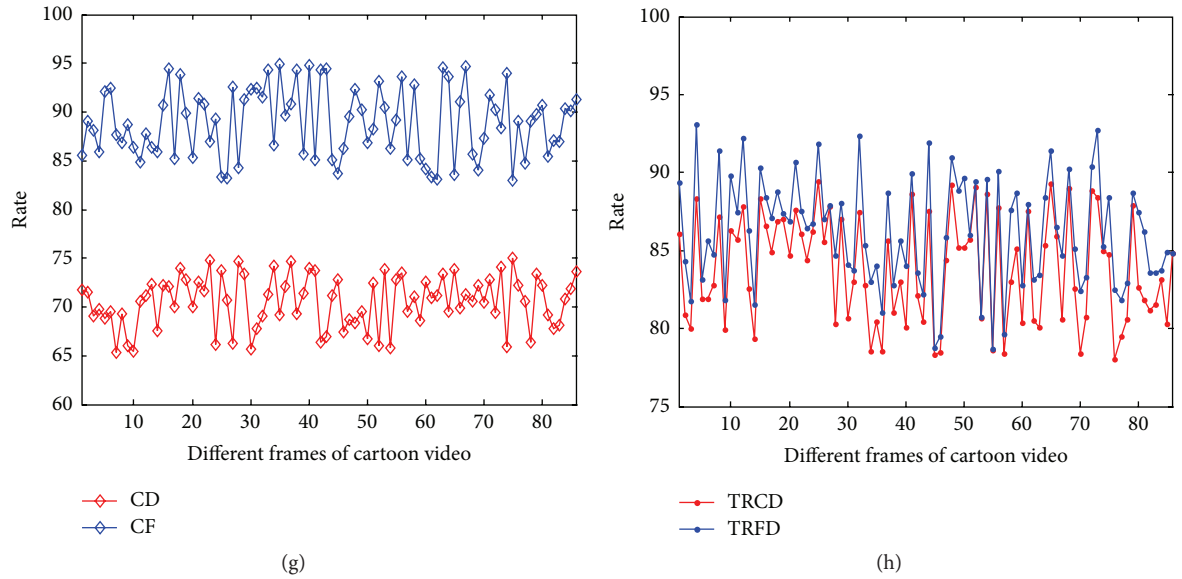


Figure 8: Algorithm details analysis results. (a), (c), (e), and (g) show corner detection rate and corner filtering rate for movie, news, sport, and cartoon, respectively. (b), (d), (f), and (h) show text regions coarse detection rate and fine detection rate for movie, news, sport, and cartoon, respectively.

For the detailed experiments of the system performance in this part, we analyzed the proposed algorithm performance step-by-step; we found that, in the whole algorithm system, corner filtering based on multiframe corner matching and text regions coarse detection under heuristics rules are two critical steps which affect the total system accuracy. In the future work, we will focus on theoretical modification of these two steps, in order to improve the performance of the whole algorithm system.

5. Conclusion

For the issue of text extraction for complex video scene, we proposed an effective video image text extraction method, which contains the coarse text region extraction based on multiframe corner matching and heuristic rules and precise text extraction under complex background based on the texture and SVM. Multiframe corner point matching was mainly used to solve Harris corner filtering problem in video image under complex background. Heuristic rules could be flexibly used to filter the candidate text regions according to the style of the video scene, which improved the efficiency of the algorithm and decreased the false alarm rate. The local image texture description was classified by SVM, and the accurate text extraction was finally achieved. The experimental results of four different video types including movie, news, sport, and cartoon, with 395 video frames, demonstrate that our method gets the average extraction accuracy of 92.7%, the average recall rate of 82.8%, and the average false alarm rate less than 6.1%, and all three indicators are obviously superior to the five comparative methods. Experimental results show the effectiveness of the proposed method for text extraction for complex video scene.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61402371 and 61461025, Natural Science Basic Research Plan in Shaanxi Province of China under Grants 2013JQ8039 and 2015JM6317, and the Fundamental Research Funds for the Central Universities under Grant 3102014JCQ01060.

References

[1] W. Wu, X. Chen, and J. Yang, "Detection of text on road signs from video," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 4, pp. 378–390, 2005.

[2] K. Jung, K. I. Kim, and A. K. Jain, "Text information extraction in images and video: a survey," *Pattern Recognition*, vol. 37, no. 5, pp. 977–997, 2004.

[3] P. Dubey, "Edge based text detection for multi-purpose application," in *Proceedings of the 8th International Conference of Signal Processing*, vol. 4, Beijing, China, November 2006.

[4] I. Ar and M. E. Karsligli, "Text region detection in digital documents images using textural features," in *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, Vienna, Austria, August 2007.

[5] W. Kim and C. Kim, "A new approach for overlay text detection and extraction from complex video scene," *IEEE Transactions on Image Processing*, vol. 18, no. 2, pp. 401–411, 2009.

[6] C. S. Shin, K. I. Kim, M. H. Park, and H. J. Kim, "Support vector machine-based text detection in digital video," in *Proceedings of*

the 10th IEEE Workshop on Neural Network for Signal Processing (NNSP '00), pp. 634–641, Sydney, Australia, December 2000.

[7] Y. Xia, Z. Ji, and Y. Zhang, "Brain MRI image segmentation based on learning local variational Gaussian mixture models," *Neurocomputing*, vol. 204, pp. 189–197, 2016.

[8] P. Shivakumara, T. Q. Phan, and C. L. Tan, "New fourier-statistical features in RGB space for video text detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1520–1532, 2010.

[9] C. Yi and Y. Tian, "Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4256–4268, 2012.

[10] X. Huang and H. Ma, "Automatic detection and localization of natural scene text in video," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR '10)*, pp. 3216–3219, Istanbul, Turkey, August 2010.

[11] L. Wu, P. Shivakumara, T. Lu, and C. L. Tan, "Text detection using delaunay triangulation in video sequence," in *Proceedings of the 11th IAPR International Workshop on Document Analysis Systems (DAS '14)*, pp. 41–45, Tours, France, April 2014.

[12] P. Shivakumara, T. Q. Phan, and C. L. Tan, "A Laplacian approach to multi-oriented text detection in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 412–419, 2011.

[13] C. Yi and Y. L. Tian, "Text string detection from natural scenes by structure-based partition and grouping," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2594–2605, 2011.

[14] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[15] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[16] Z. Li, G. Liu, X. Qian, D. Guo, and H. Jiang, "Effective and efficient video text extraction using key text points," *IET Image Processing*, vol. 5, no. 8, pp. 671–683, 2011.

[17] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[18] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, "Video OCR for digital news archive," in *Proceedings of the IEEE International Workshop on Content-Based Access of Image and Video Database*, pp. 52–60, Bombay, India, January 1998.

[19] M. R. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 243–255, 2005.

[20] H. X. Song, N. N. Zhao, and X. H. Xu, "Extraction of text under complex background using wavelet transform and support vector machine," in *Proceedings of the IEEE International Conference on Mechatronics and Automation (ICMA '06)*, pp. 1493–1497, Luoyang, China, June 2006.

[21] P. Shivakumara, T. Q. Phan, and C. L. Tan, "A robust wavelet transform based technique for video text detection," in *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR '09)*, pp. 1285–1289, Barcelona, Spain, July 2009.