Hindawi
Mobile Information Systems
Volume 2021, Article ID 7376572, 8 pages
https://doi.org/10.1155/2021/7376572

Research Article

Adaptive Video Caption Detection and Extraction Method Based on Color Filtering Principle

Peng Ren, Genlin Zhao, and Yongjun Liu

Changshu Institute of Technology, Jiangsu, Suzhou 215500, China

Correspondence should be addressed to Yongjun Liu; lyj@cslg.edu.cn

Received 30 October 2021; Accepted 29 November 2021; Published 13 December 2021

Academic Editor: Sang-Bing Tsai

Based on the principle of color (RGB) filtering, an improved adaptive video caption detection and extraction method is proposed. Firstly, the principle of the color filtering algorithm used in the video caption detection and extraction method is analyzed, and then the algorithm is improved adaptively according to the caption pixel size to filter the noise. Finally, experiments verify the effect of this method in extracting subtitles from video. The experimental results show that the accuracy of detecting and extracting subtitles in color video is as high as 99.3%.

1. Introduction

At present, video media forms appear in all aspects of the Internet; people's study, work, and life highly depend on video. The relevant algorithms of video caption extraction have high computational complexity and a huge amount of video data, especially the rapid growth of video data in the Internet environment, which puts forward higher and higher requirements for the performance of video caption detection and extraction [1]. Video captions are usually divided into two types [2]: one is the text information carried by the object itself or the environment in the video such as the text on books and advertising boards, license plates, and pedestrians' clothes which are collectively called background captions (scene captions) and the other kind of text captions is artificially added to videos, which are unrelated to the scene information in videos and are generally called "hard captions." Hard captions, like watermarks, are combined with video images, which cannot be simply separated and cannot be directly analyzed [3]. This paper studies the method of extracting hard caption detection by color filtering principle.

In essence, the research on video caption detection and extraction is to study how to detect, locate, and segment the caption in video effectively and quickly and form binary images. At present, the detection of video captions extract mainly has the following four methods: Sato et al. proposed to detect captions based on the edge of the video frame [4]. First, the information is detected by the edge feature of the video captions, and then the areas outside the captions are removed according to conditions such as the size and density of the captions and then projected in the horizontal and vertical directions, respectively. The captions are finally determined according to the projection. This method can quickly detect captions, but its disadvantage is that the threshold needs to be debugged several times before detection to improve the success rate. Threshold is not applicable to other videos, and the error rate is high when the background is complex. Kim et al. proposed texture analysis-based caption detection [5], which determines whether a certain pixel point or pixel block belongs to text based on image texture. The advantage of this method is that it can effectively reduce the error rate in complex background, but it also has the disadvantage of low efficiency caused by the complexity of the algorithm and the large amount of calculation and the inaccurate positioning of the caption area. Lienhart and Wernicke proposed machine learning-based caption detection to construct a machine learning machine to classify image subblocks by learning support vector machines and neural networks [6]. The detection error rate

of this method is low, but the calculation is quite complex, and the detection effect is affected by the samples of the previous training, so the deviation is easy to occur. The connected domain-based caption detection method proposed by Jiang et al. uses the Niblack method to decompose images into many connected components and then uses a two-stage classification module consisting of a cascade classifier and a SVM classifier to verify the text features of the connected components [7]. The design of this method is complicated, although the accuracy is relatively high, but it is difficult for the detection and processing of large quantities of video captions.

Although the above methods have high detection accuracy, they also have their own shortcomings. There is still a large research space to further improve the accuracy. Nowadays, videos are usually in color, and the captions are mostly white, with a few other colors. The adaptive color filtering algorithm adopted in this paper focuses on the second correction after color filtering, which has precision, directivity, and exclusivity. Compared with other methods, it can improve the overall detection of captions and suppress the noise points around the text. Especially in the movie video, the accuracy of white caption detection on black background is as high as 99.3%. The effectiveness of this method in video caption extraction is verified by experiments.

2. The Traditional Color Filtering Algorithm

A video or a series of video captions often use uniform color based on the characteristics through the color filter using clustering and some other methods. The caption from video successfully in the background has good detection effect of uniform color, but when the captions have different color, it needs to set up testing for many times. It has practical operation significance to design an algorithm based on color filtering principle to better detect and extract video caption text. The essence of color filtering is to emphasize the main color of captions while weakening other background video colors, that is, to take one of RGB three channel values or superimpose normalized gray values as color features.

2.1. Projection of RGB Cube on Hexagon. R (red), G (green), and B (blue) are the three primary colors of light, and the three primary colors are the simplest form of color. Most colors in nature can be obtained by mixing these three primary colors in a certain proportion; conversely, any kind of color can be decomposed for these 3 kinds of primary colors. Generally, R, G, and B are represented by three values in the range [0, 255], representing a color in 3D space, and the three values R, G, and B correspond to X, Y, and Z values in the coordinate system. The RGB cube is tilted 45° on the x axis and then 35.264° on the Y axis. After the second tilt, the black vertex is at the bottom, and the white vertex is at the top, and they are both on the Z axis. When looking down at the cube from the top, a hexagonal appearance projection with chromaticness (hue) order is obtained. A hexagon of the same size is drawn as the top view of the cube. All the angles of the hexagon correspond to the angles of the cube, and the colors should also correspond. The white apex corner of the cube is projected onto the center of the hexagon, and the black is omitted. If each color is mapped to a hexagon, a standard full color hexagon is obtained. The position of points with RGB value of coordinates (51, 153, 204) on the hexagon is shown in Figure 1.

2.2. Gray Processing Algorithm of Color Filtering. When the background color of caption is pure color or translucent strip, the algorithm can detect and extract text by using the method of connected domain analysis [8]. Such captions can be generated as follows:

$$D(i\ ,\ j) = \alpha A(i, j) + (1 - \alpha)\, B(i, j) + C(i, j), \quad (1)$$

where $A(i, j)$, $B(i, j)$, $C(i, j)$, and $D(i, j)$, respectively, represent the pixel value of the original video image, the pixel value of the translucent strip (solid color) background, the pixel value of the caption, and the pixel value of the final image at (i, j).

When the video caption background is complex, the color image needs to be converted into gray image first. Each pixel in a color image is determined by R, G, and B components, and each component is valued at [0, 255]. Therefore, the range of pixels varies greatly. In gray image, R, G, and B components are taken to the same value, so the range of each pixel point is 255, white is 255, and black is 0. Grayscale image has less computation than color image, and grayscale image can also reflect the overall and local color and luminosity distribution of the image. There are four commonly used methods for image grayscale [9]: component method, average method, maximum method, and weighted average method.

2.2.1. Component Method. The luminosity of one of R, G, and B components in the color video image is taken as the gray value of the image. Taking R channel as an example, the formula can be expressed as follows:

$$\text{gray}\, R(i, j) = R(i, j). \quad (2)$$

2.2.2. Average Value Method. The average luminosity value of R, G, and B components in the color video image is taken as the gray value, and the formula is expressed as follows:

$$\text{gray}\,(i, j) = \frac{(R(i, j) + G(i, j) + B(i, j))}{3}. \quad (3)$$

2.2.3. Maximum Value Method. The maximum luminosity of R, G, and B components in the color video image is taken as the gray value of the gray graph, and the formula is expressed as follows:

$$\text{gray}\,(i, j) = \max\{R(i, j), G(i, j), B(i, j)\}. \quad (4)$$

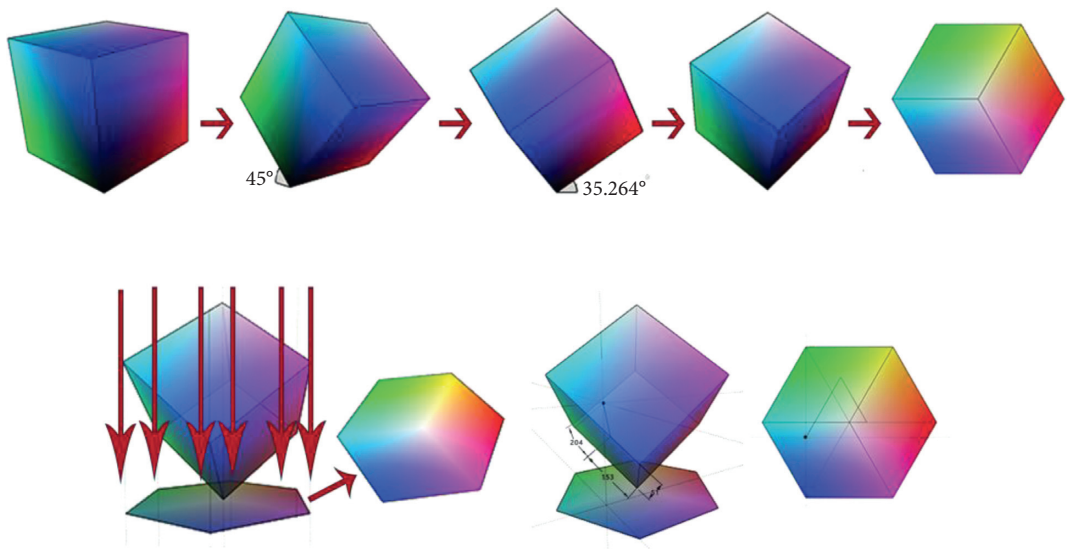

Figure 1: Hexagons after RGB cube projection.

2.2.4. Weighted Average Method. The R, G, and B components of the color video image are weighted and averaged with different weights. Among them, human eyes are more sensitive to green and less sensitive to blue. The formula is expressed as follows:

$$\text{gray}(i, j) = 0.299 * R(i, j) + 0.578 * G(i, j) + 0.114B(i, j). \quad (5)$$

It can be seen from the traditional methods of image graying that the first three algorithms are all based on luminosity value. Although they simplify the calculation process and increase the calculation speed, they have good effects when the picture and captions are in different color systems, but when they are in similar color systems, large detection errors will be generated. The last method adopts the weighted average method, which simplifies the human eye simulation into a proportional calculation method. Although the detection effect of captions is improved, the uniqueness of the weighted average results in more and larger noise points, which cause serious interference to the later recognition.

3. The Realization of Adaptive Color Filtering Algorithm

The traditional color filtering is changing to gray scale based on the image luminosity, for example, component method, average method, maximum method, and weighted average method. These methods have a fixed formula, that is, R, G, and B have a fixed value. Adaptive the color filtering algorithm is proposed in this paper which will first convert RGB to H (chroma) value, and then, according to the chromaticity threshold of the sample, the particle features around the detected caption text are analyzed and processed, and the secondary correction is carried out in the way of block elimination, so as to realize the function of strengthening the sample chromaticity caption text to suppress interference factors, and finally, the detection and extraction of video caption are completed efficiently.

3.1. RGB Model Conversion. Conversion should follow certain principles. When the RGB gap between the color of the video and the color of the selected caption is larger than the set value, it needs to be filtered. When the color difference between the video color and the color of the selected caption is greater than the set value, it needs to be filtered. When the luminosity of the video color is lower than the set value, it needs to be filtered. When the luminosity of video color is higher than the set value, it needs to be filtered. Minimum saturation: when the color saturation of the movie is lower than the set value, it needs to be filtered. When the color saturation of the movie is higher than the set value, it needs to be filtered. The value range of major items is shown in Table 1.

Among them, the RGB gap is approximately the definition of subtitles. The higher the value is selected during filtering, the clearer it is. However, the higher the value, the noise interference will appear in nonsubtitles, which will affect the extraction effect in the next step. When using HSL to filter, the order of judgment is as follows: lowest color-$\longrightarrow$ highest color $\longrightarrow$ lowest luminosity $\longrightarrow$ highest luminosity$\longrightarrow$ RGB gap $\longrightarrow$ hue gap.

Before performing the calculation, chroma H is defined as follows. Chroma is roughly the angle between the vector and a point in the projection, and red is 0°. Hue H' is the distance from the point to the edge of the hexagon. The conversion formula between chroma H and RGB model is divided into four sections:

(1) When $C = 0$, $H' = \text{undefined}$

(2) When $M = R$, $H' = (G - B/C)\text{mod}6$ and $H = 60^\circ \times H'$

(3) When $M = G$, $H' = (B - R/C) + 2$ and $H = 60^\circ \times H'$

Table 1: Value range of major items in the RGB model.

Major items	Value range
RGB difference	[0–255]
Hue difference	[0–181]
Luminosity	[0–100]
Chroma	[0–100]

(4) When $M = B$, $H' = (R - G/C) + 4$ and $H = 60^\circ \times H'$

The formula is shown as follows:

$$H' = \begin{cases} \text{undefined}, & \text{if } C = 0, \\ \dfrac{G-B}{C} \bmod 6, & \text{if } M = R, \\ \dfrac{B-R}{C} + 2, & \text{if } M = G, \\ \dfrac{R-G}{C} + 4, & \text{if } M = B, \end{cases} \tag{6}$$

$$H = 60^\circ \times H'$$

As shown in Figure 2, the coordinate position on the hexagon (51, 153, 204) is detected. First, the values of R, G, and B are normalized to the range of [0, 1], as shown in the following calculation process:

$$\begin{aligned} R &= \frac{R}{255}, R = \frac{51}{255} = 0.2, \\ G &= \frac{G}{255}, G = \frac{153}{255} = 0.6, \\ B &= \frac{B}{255}, B = \frac{204}{255} = 0.8. \end{aligned} \tag{7}$$

Max and min values of R, G, and B are as follows:

$$\begin{aligned} M &= \max(R, G, B), M = \max(0.2, 0.6, 0.8) = 0.8, \\ m &= \min(R, G, B), m = \min(0.2, 0.6, 0.8) = 0.2. \end{aligned} \tag{8}$$

The color saturation value C (chroma) on the hexagon is then calculated and defined as the distance from the point to the origin. The color saturation here is the relative size of the hexagon through the point:

$$\begin{aligned} C &= \frac{\mathrm{OP}}{\mathrm{OP}'}, \\ C &= M - m, \\ C &= 0.8 - 0.2 = 0.6. \end{aligned} \tag{9}$$

A condition check is performed after R, G, and B, and C values are determined. For the coordinates (51, 153, 204), since $M = B$, $H' = ((r - g)/C) + 4$ will be used. Checking the hexagon again, $(R - G)/C$ is the length of the BP line segment.

$$\text{segment} = \frac{(R-G)}{C} = \frac{(0.2 - 0.6)}{0.6} = -0.6666666666666666. \tag{10}$$

The line segment on the inner hexagon is placed, which is outset as R (red) at 0°. If the length of the line segment is positive, it should be on RY; if it is negative, it should be on RM.

Because it is −0.666666666666, it is on the edge of RM.

As shown in Figure 3, since $M = B$, P points to B when moving the position of the segment. The color of blue is at 240° in the figure, hexagon has 6 sides, each side corresponds to 60°, 240/60 = 4, and P needs to be moved (increased) 4 times (since it is 240°). The end position after moving is located at P position, and the length of RYGCP is as follows:

$$\begin{aligned} \text{segment} &= \frac{(R-G)}{C} = \frac{(0.2 - 0.6)}{0.6} = -0.6666666666666666, \\ \text{shift} &= 4, \\ \text{RYGCP} &= \text{segment} + \text{shift} = 3.3333333333333335. \end{aligned} \tag{11}$$

The circumference of the hexagon is equal to 6, which corresponds to 360 degrees. The distance from coordinates (53, 151, 204) to 0° is 3.333333333335. If we multiply 3.333333333335 by 60, we will get the position in terms of angles.

$$\begin{aligned} H' &= 3.3333333333333335, \\ H &= H' * 60 = 200^\circ. \end{aligned} \tag{12}$$

In the case of $M = R$, since one end of the line segment is placed at R (0°), when the length of the line segment is positive, there is no need to move the line segment to R, and the position of P is positive. When the length of the line segment is negative, it needs to be moved 6 times because a negative value indicates that the angular position is greater than 180°, and a complete rotation is required. Therefore, regardless of the scheme hue (1) = $(g - b)/c$ or scheme hue (2) = $((g - b)/c) * 60$, it is not applicable to the negative line segment length. This scheme uses hue = $((g - b)/c) + ((g - b)/c) * 60$, which is correct for both negative and positive values.

3.2. Adaptive Color Filtering. In the process of color filtering, blocky background images with the same color will be detected together with caption text, resulting in block effect around the text [10]. In general, the smoother the image, the more obvious the block effect. In order to reduce the block effect and improve the image quality, this paper improves the average block effect removal method in adjacent images. The pixel size of strokes of caption text in the video is generally less than 9×9, which means that color blocks of pixels above 9×9 should be filtered. L and R represent horizontal adjacent subimages to be filtered, as shown in Figure 4. There will be a block effect at the adjacent boundary of subimage L and R. In order to remove the block effect of

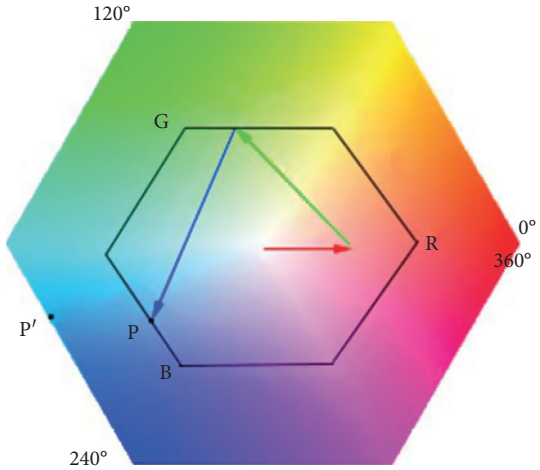

Figure 2: The position of the coordinates (51, 153, 204) on the hexagon.

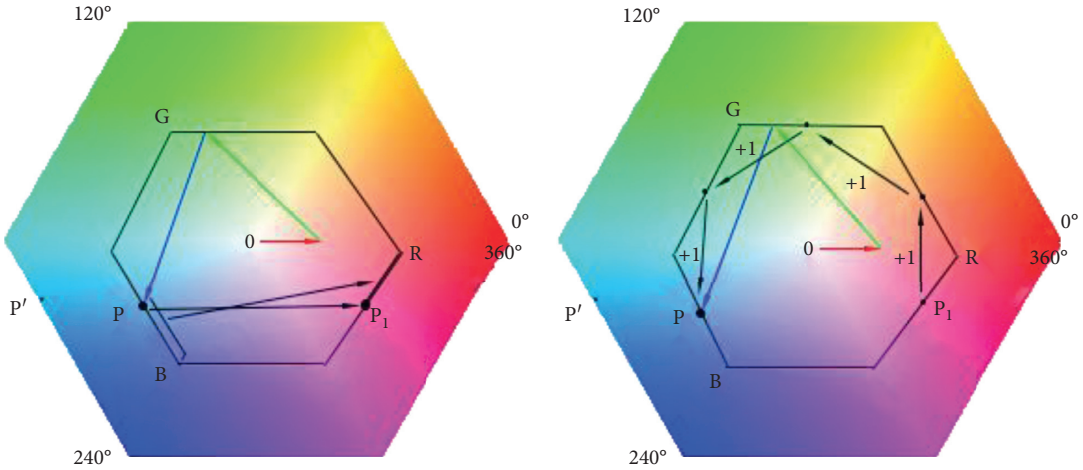

Figure 3: Moving image for calculating the H value of the coordinate point.

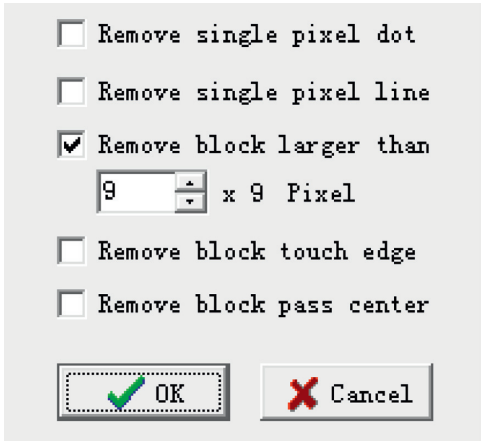

Figure 4: Removal of adaptive blocks for color filtering.

horizontal adjacent subimages, mean filtering operation is performed on the boundary pixels adjacent to L and R, and the algorithm formula is expressed as the following equations:

$$L(m,n) = \frac{\sum_{i=0}^{n_0} L(m,n-i) + \sum_{i=0}^{n_0-1} R(m,i)}{N}, \quad 0 \le m \le 9, n = 9, \tag{13}$$

$$R(m,n) = \frac{\sum_{i=0}^{n_0-1} L(m,9-i) + \sum_{i=0}^{n_0} R(m,n+i)}{N}, \quad 0 \le m \le 9, n = 0. \tag{14}$$

The template size is $1 \times N, N = 2n_0 + 1$, n_0 is an integer, and $1 \le n_0 \le 9$.

U and D represent the adjacent subimages filtered by local homomorphism in vertical direction. Block effect will appear at the adjacent boundary of subimage U and D. In order to remove the block effect of vertically adjacent subimages, mean filtering is carried out on the boundary pixels adjacent to U and D, and the algorithm formula is expressed as the following equations:

$$U(m,n) = \frac{\sum_{i=0}^{M_0} U(m-i,n) + \sum_{i=0}^{m_0-1} D(i,n)}{M}, \quad 0 \le m \le 9, n = 9, \tag{15}$$

$$D(m,n) = \frac{\sum_{i=0}^{M_0-1} U(7-i,n) + \sum_{i=0}^{m_0} D(m+i,n)}{N}, \quad m = 0, 0 \le n \le 9. \tag{16}$$

The template size is $M \times 1, M = 2m_0 + 1, m_0$ is an integer, and $1 \le m_0 \le 9$.

As shown in Figure 5, if color block removal is not carried out, there will be a small number of dispersive parts of the same color after color filtering, and these large spots (blocks) will form interference factors near the caption text, which will often be misrecognized and misread in the recognition after detection. The adaptive color filtering scheme proposed by block removed to a certain size with color interference speckles delete has good effect, and repeated experiments show that compared with the traditional color filtering detection, the adaptive scheme proposed in this paper can greatly improve the detection accuracy. The accuracy is improved about 10%. This method achieves the purpose of reducing block (noise) interference without affecting the detected subtitle text and clears the obstacles for further recognition.

4. Analysis of Experimental Results

In order to test the proposed method, caption detection experiments are carried out on videos of news, movies, and TV programs. Each type of video material in the experiment includes one or two colors of captions. The length of selected videos is more than 15 minutes, and the total number of captions is more than 1000. Figure 6 shows the screenshots of film caption detection in the experiment of three types of video caption detection. The detection accuracy of different types of video is calculated through experiments:

$$\text{accuracy} = \frac{\text{number of correctly detected captions}}{\text{total number of captions in a video}} \times 100\%. \tag{17}$$

Due to the presence of two color captions in the TV program video in the experiment, the detection accuracy was only 92.2% lower than the other two types in one detection process. As shown in Figure 7, the accuracy of detecting the main color is guaranteed, with only a small amount of tiny pixel noise, but the subtitles of the second color other than the main color appear blank or other color blocks during filtering, and the subtitles in the red box need to be detected twice to further improve the accuracy.

News video compared with TV video background environment is similar, but the news video captions have more uniform color; accuracy can reach 94.5%, as shown in Figure 8; when part of the video background and the caption show the same color system, it will interfere with the detection. During adaptive color filtering, the detection accuracy will decline, small noise will appear locally, and a small number of repeated sentences or broken sentences will be left.

Since the caption background is solid color and the picture is full in color, there are few interference factors in caption detection. Except for the difficulty in the detection of a few dialogues and repeated sentences, all the others can be accurately detected without any noise. As shown in Figure 9, the detected caption text has clear edges and the picture has no noise. In the experiment, the accuracy of adaptive color filtering in the detection of captions in the film is up to 99.3%, and the repeated dialogue captions in the video can be accurately judged and detected separately.

The experiment compares the detection accuracy with the traditional color filtering method and other three methods [11], namely, the method of detecting caption based on frame difference, the method of detecting caption based on edge detection, and the method of detecting caption based on color

Figure 5: Comparison of original video and adaptive block before and after removal.

Figure 6: Subtitle detection of the film "pride and prejudice."

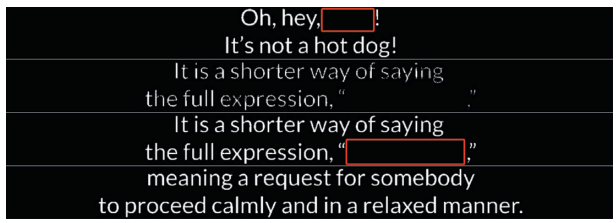

Figure 7: Subtitle detection results of TV program "learn English with SpongeBob SquarePants."

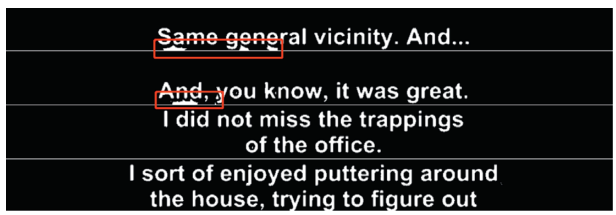

Figure 8: Image of news video captions detection results.

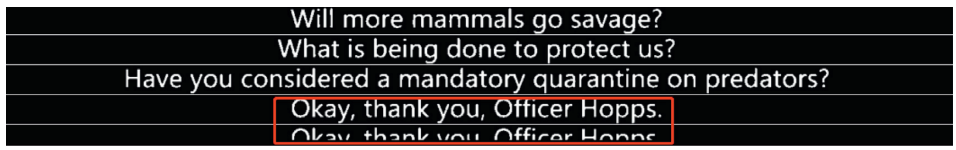

Figure 9: Subtitle detection results of movie (pride and prejudice).

clustering. According to the comparison of the accuracy of caption detection of the four schemes in Table 2, the method studied in this paper has a higher accuracy in detecting caption in extracted videos, especially in detecting caption of videos with a solid color background and a single color caption. Therefore, it can be concluded that the adaptive video caption detection and extraction method based on the color filtering principle in this paper achieves improved accuracy compared with the traditional color filtering, and the detection effect is better than the other three methods.

Table 2: The accuracy of caption detection in four schemes (unit: %).

Caption type	Scheme Accuracy				
	Frame difference	Edge detection	Color clustering	Traditional color filter	Adaptive color filter
News	88.2	85.2	81.2	80.3	94.5
Film	91.0	86.9	83.1	91.5	99.3
TV program	87.5	86.5	80.5	77.8	92.2

5. Conclusion

The adaptive video caption detection and extraction method based on the principle of color filtering can detect and extract the video caption text more ideally and has great advantages in both accuracy and speed. The video frame images in the experiment in this paper are all 1920×1080 HD videos. There are no experiments on low-definition videos and higher-definition videos and no detection and extraction experiments on scrolling captions, pop-frame captions, color-changing captions, and other multitype captions, all of which need to be compared with data in further experiments. At the same time, it should be noted that the method of adaptive color filtering for caption detection also needs to be improved, and the next research work will continue to improve the algorithm to further improve the accuracy of caption detection in complex background videos.

Data Availability

The data of each type of video material used to support the findings of this study are currently under embargo while the research findings are commercialized. Requests for data, after publication of this article (6/12 months), will be considered by the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper is supported by the Philosophy and Social Sciences Research Project of Jiangsu Universities (2020sja1406).

References

[1] F. Gao and Y. Liu, "Acceleration of video caption retrieval algorithm based on Intel MIC," *Computer Engineering and Science*, vol. 37, no. 4, pp. 634–640, 2015.

[2] Y. Wang, J. Yan, and H. Zheng, "An adaptive method for detecting and location text in vedio frame," *Journal of Computer Applications*, vol. 24, no. 1, pp. 134-135, 2004.

[3] Q. Wang and L. Q. Chen, "Extraction of caption in videos," *Computer Engineering and Applications*, vol. 48, no. 5, pp. 177-178, 2012.

[4] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, "Video OCR for digital news archive," in *Proceedings of the1998 IEEE International Workshop on Content-Based Access of Image and Video Database*, pp. 52–60, Bombay, India, 1998.

[5] K. I. Kim, K. Jung, and J. H. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1631–1639, 2003.

[6] R. Lienhart and A. Wernicke, "Localizing and segmenting text in images and videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 4, pp. 256–268, 2002.

[7] R. J. Jiang, F. H. Qi, L. Xu, and G. R. Wu, "Using connected-components features to detect and segment text," *Journal of Image and Graphics*, vol. 11, no. 11, pp. 1653–1656, 2006.

[8] W. S. Zhang, "Research on TV video caption recognition and retrieval technique," *Beijing University of Posts and Telecommunications*, Mater thesis, Beijing, China, 2016.

[9] J. Shangguan, "Video caption localization and recognition," *Xiamen University*, Mater thesis, Xiamen, China, 2017.

[10] Y. F. Zhang and M. H. Xie, "Color image enhancement algorithm based on HSI and local homomorphic filtering," *Computer Applications and Software*, vol. 30, no. 12, pp. 303–307, 2013.

[11] Z. H. Wang, J. T. Li, S. Y. Xie, J. Zhou, H. J. Li, and X. Fan, "Two-stage method for video caption detection and extraction," *Computer Science*, vol. 45, no. 8, pp. 50–53 + 62, 2018.