Multimedia Tools and Applications (2024) 83:5437–5450
https://doi.org/10.1007/s11042-023-15829-5

Key frame extraction method for lecture videos based on spatio-temporal subtitles

Yunzuo Zhang[1] · Yi Li[1] · Zhaoquan Cai[2] · Xuejun Wang[1] · Jiayu Zhang[1] · Shui Lam[3]

Received: 10 January 2022 / Revised: 23 February 2023 / Accepted: 10 May 2023 /
Published online: 2 June 2023

Abstract
Affected by the Corona Virus Disease 2019 (COVID-19), online lecture videos have witnessed an explosive growth. In the face of massive videos, this paper proposes a method for extracting key frames of lecture videos based on spatio-temporal subtitles, which can efficiently and quickly obtain effective information. Firstly, the spatio-temporal slices of subtitle area of the video sequence are extracted and spliced along the time axis to construct the video spatio-temporal subtitle. Then, the video spatio-temporal subtitle is processed in binarization, and the projection method is used to construct the SSPA curve of the video spatio-temporal subtitle. Finally, a selection method for steady-state key frame is designed, that is, the key frame extraction is realized by combining curve edge detection and subtitle existence threshold, which ensures the robustness of the proposed method. The test results of 8 videos show that the average value of the comprehensive index F_1-score of the key frame extracted by the algorithm can reach 0.97, the average precision is 0.97, and the average recall rate is 0.98. It can effectively extract the key frames in lecture videos, and compared with other algorithms, the average running time is reduced to 0.072 of the original, which is helpful to extract video information quickly and accurately.

Keywords Lecture video · Spatio-temporal subtitle · Key frame extraction · Steady-state key frame

1 Introduction

Since the massive spread of COVID-19 in December 2019, the virus has spreads more than 200 countries and regions around the world, which not only poses a major threat to the lives and health of people around the world, but also affects people's way of production and life, catalyzing the development and maturation of online offices,

✉ Yunzuo Zhang
zhangyunzuo888@sina.com

[1] Shijiazhuang Tiedao University, Shijiazhuang, People's Republic of China

[2] Shanwei Institute of Technology, Shanwei, People's Republic of China

[3] California State University, Long Beach, CA, USA

information technology in education and networking of resources [1, 16]. With the normal development of the prevention and control of the COVID-19 epidemic, we have entered the Post-epidemic Era, online education and teaching, and online academic conferences have become a major highlight of the ongoing presence now [4]. People can easily use a variety of online videos to learn professional knowledge, thanks to the popularity of internet technology. Academic lecture videos have significant advantages and are rapidly developing, as they are not restricted by time and space. How to efficiently obtain effective information of interest from numerous web videos has become an urgent problem to be solved at present [5, 25, 27]. Handling massive amounts of video is no longer possible through traditional manual retrieval methods such as fast forward, pause and rewind. The video key frame extraction technology has been developed to overcome the problem of large amounts of video data and low retrieval efficiency [33]. The key frame extraction technique aims to extract key and meaningful series of still images from the original video to express the semantic information of the video as completely as possible [29]. Therefore, the effect of key frame extraction directly affects the performance of video retrieval and people's visual experience.

The existing key frame extraction techniques are usually based on shot segmentation and rely on video shot detection techniques, which segment video into multiple shots using the visual differences between frames at the boundaries of the shots [8, 14]. Zhang et al. [30] used thresholding to identify shots in the video data and then selected the key frame in each shot. Qu et al. [18] used color features for shot detection and selected the frame with the highest image entropy value as the key frame within the shot. The key frame extracted by this type of method can effectively consider the content of the shot, but there is a high dependency on the threshold value during the shot segmentation process. Lo et al. [15] obtained inter-frame difference vectors by the histogram method and divided the inter-frame differences into two categories with and without shot transitions using the clustering method. The clustering method eliminates the dependence of the shot boundary detection on thresholds, but requires setting the classes of clusters in advance. Bai et al. [3] proposed a key frame extraction method based on an improved clustering method. The method treats the initial result of the primary hierarchical clustering algorithm as the initial condition for the secondary artificial immune clustering algorithm and obtains key frames. The method prevents the disadvantage of manually setting the clustering centers and numbers, but lacks information on the temporal order and dynamics between image frames. Prabavathy et al. [12] proposed a histogram difference method based on fuzzy rules for shots boundary detection, which overcomes the disadvantages of threshold-based components but was prone to losing spatial information. Most of the existing key frame extraction techniques analyze the underlying features of the video, which leads to the drawback that the extraction results cannot represent the true content of the video accurately and comprehensively [6, 10]. In the face of different application scenarios, the existing technologies are characterized by the unique picture information. Wang et al. [23] proposed a template and color moment based shot detection method of news video presenter to extract presenter frames as key frames of news videos to generate summaries. Wu [26] proposed a slow playback shot detection algorithm to extract the highlight slow playback shot as the key frame of the sports video to establish the summary. Zhang et al. [31] proposed the key frame extraction algorithm based on frequency domain analysis to extract the frame with the maximum local center offset of the moving object as the key frame of the surveillance video.

The investigation shows that the embedded text of video has great research value and it is a significant clue to understand the content of video [32]. In addition, video subtitles mostly appear below the video, which is in sharp contrast with the background of subtitles. In general, the subtitle is in sharp contrast with the background, and information of it is concise and comprehensive, which has a good summary effect on the video content [2]. Jin et al. [11] proposed that the subtitles carried by video data are the core of extracting the high-level semantics of the video. Motivated by this consideration, this paper takes the subtitle of the lecture video as the characteristic to propose a key frame extraction method based on spatio-temporal subtitle. Initially, a video spatio-temporal subtitle is built by using the spatio-temporal slice. Then, the projection method is used to construct the pixels accumulation curve of the spatio-temporal subtitle (SSPA). Finally, a steady-state key frame selection method for video is proposed by combining curve edge detection and subtitle existence threshold. Unlike the key frame extraction method in the deep learning perspective [20, 22], which has the defects of complex feature extraction, difficult algorithm framework and large computational amount [7, 9, 13, 21], the proposed method adopts the spatio-temporal slice technique in the video feature extraction process to achieve the local operation, which greatly reduces the computational amount. Therefore, the method of this paper is helpful to achieve accurate and fast extraction of the lecture video key frames and it is interesting and worthwhile to research.

The contributions of this paper are fivefold:

(1) We develop a video spatio-temporal subtitle construction method using the spatio-temporal slice technique.
(2) We built the pixels accumulation curve of the spatio-temporal subtitle (SSPA) to record the change of the video subtitle existence status.
(3) We propose a key frame extraction method of lecture video based on the video spatio-temporal subtitle.
(4) We adopt a method for selecting steady-state key frame using curve edge detection and subtitle existence thresholds.
(5) We compared the proposed key frame extraction method with existing methods on the public video segments to demonstrate the effectiveness of this method.

The remainder of this paper is organized as follows. Section 2 constructs the spatio-temporal subtitle for video and verifies its validity. Section 3 illustrates the proposed steady-state key frame extraction method based on the video spatio-temporal subtitle. Section 4 presents the experimental results and the experimental comparative analysis with the existing methods. Finally, conclusions and future work prospects are provided in Sect. 5.

2 Video spatio-temporal subtitle

In this paper, key frames are defined as video frames carrying subtitles that are in a steady state. Using the traditional inter-frame difference-based method [24] to extract the frames with subtitle is feasible, but this method unfolds the analysis based on the global image, which leads to a computationally intensive and very time consuming algorithm, making it difficult to meet the needs of efficient video browsing. Therefore, in this paper, we analyze video spatio-temporal subtitle to detect changes in video subtitles

Fig. 1 3D image sequence representation of the video

and then extract the key frames. The spatio-temporal subtitles of the video are obtained by sampling spatio-temporal slices [17] of the video, that is, by sampling video images that are unfolded in chronological order on a timeline at the same locations where captions exist. For example, by extracting a row or a column of pixels from a sequence of images and combining them to form a two-dimensional image, where one dimension is time t and the other dimension is the direction x or y, representing a horizontal slice in the $x-t$ dimension and a vertical slice [19] in the $y-t$ dimension, respectively. Figure 1 shows the 3D image sequence of the video, which denoted by $V(x, y, t)$. Figure 2 illustrates the schematic of horizontal spatio-temporal slice.

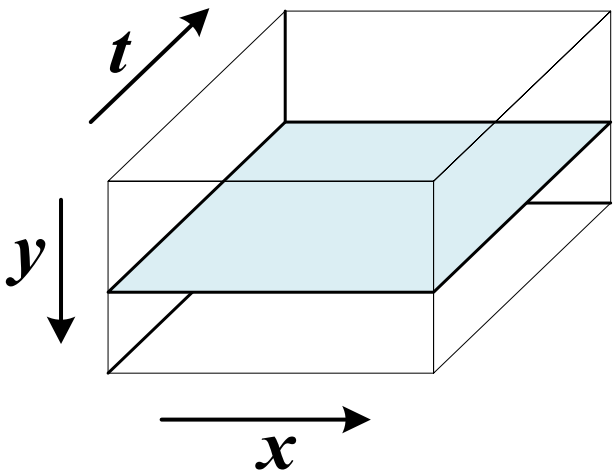

Fig. 2 Horizontal spatio-temporal slice

The spatio-temporal subtitles of $V(x, y, t)$ is shown in Eq. (1).

$$S = \begin{bmatrix} p_1^1 & p_2^1 & \cdots & p_{i-1}^1 & p_i^1 & \cdots & p_{L-1}^1 & p_L^1 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ p_1^{1-1} & p_2^{j-1} & \cdots & p_{i-1}^{j-1} & p_i^{j-1} & \cdots & p_{L-1}^{j-1} & \mathrm{p}_L^{j-1} \\ p_1^j & p_2^j & \cdots & p_{i-1}^j & p_i^j & \cdots & p_{L-1}^j & p_L^j \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ p_1^W & p_2^W & \cdots & p_{i-1}^W & p_{\mathrm{i}}^W & \cdots & p_{L-1}^W & p_L^W \end{bmatrix} \tag{1}$$

Among them, $S\left(p_i^j\right)$ represents the pixels of the video frame of $V(x, y, t)$ at a specific position in the subtitle area, where $x = j$, $t = i$ and y is a known value. The following conditions are satisfied, where $j \in [1, W]$, $i \in [1, L]$, in addition, W is the width of the video frame and L is the length of the video stream.

According to Eq. (1), the spatio-temporal subtitle extracts only one row of pixels in the subtitle image, thus preserving the complete time-domain information of the video, while the lack of space domain information has little impact on the detection of video caption changes. The spatio-temporal subtitle has the advantages of low computational effort and high interference resistance. Figure 3 shows an example of video spatio-temporal subtitles.

The example of spatio-temporal subtitles for a video is shown in Fig. 3, where a representative row is cut for each frame. The horizontal representation is the video time domain information, which is the length of the video; the vertical representation is the video space information, which is the width of the video frame. As shown from Fig. 3 that in the video spatio-temporal subtitles, the area without subtitles is pure black, and the subtitled area is displayed in white. The information such as the duration and length of the subtitles are visible, and the length and texture of different subtitles have distinct characteristics. It can be seen that it is feasible to use video spatio-temporal subtitles to detect the moment of change of video subtitles.

Fig. 3 Example of the spatio-temporal subtitles for a video

3 Video key frame extraction based on spatio-temporal subtitle analysis

The subtitles stay of a lecture video usually lasts for more than a few seconds. The same subtitles reflect the same video content, and the moments when the subtitles alternate (i.e. appear and disappear) attract the most visual attention. Based on this observation, in order to ensure the stability of the extracted frames with subtitles, this paper selects the video frame at the middle position between the appearance and disappearance of the subtitle as the key frame. The SSPA values of the video can accurately reflect the changes of the video subtitles. Therefore, this paper is based on the spatio-temporal subtitle of the video to carry out the analysis. Firstly, a video spatio-temporal subtitle is built by using the spatio-temporal slice. Then, the projection method is used to construct the pixels accumulation curve of the spatio-temporal subtitle (SSPA). Finally, a steady-state key frame selection method for video is proposed by combining curve edge detection and subtitle existence threshold.

Based on the above analysis and the example presented in Sect. 2, it is clear that the video spatio-temporal subtitles image has obvious segmentation lines, which means that the changing state of the video subtitle presence can be clearly captured. Therefore, this section proposes a key frame extraction method based on saptio-temporal subtitle analysis. Figure 4 illustrates the framework of proposed method.

The details of these steps will be discussed as following.

(1) SSPA curve Establishment

After reading the input video and decomposing it into single frames, the spatio-temporal subtitles represented by S is obtained from the video sequence according to Eq. (1). The brightness of the pixel in the spatio-temporal subtitles characterizes the relative saliency of the subtitles, and the stronger the saliency, the higher the accumulated pixel value. Based on Eq. (1), the SSPA of the i^{th} frame in video $V(x, y, t)$ can be calculated by Eq. (2).

$$\text{SSPA}(i) = \sum_{j=1}^{W} P\left[S\left(p_i^j\right)\right] \tag{2}$$

Among them, $P\left[S\left(p_i^j\right)\right]$ can be expressed by Eq. (3).

$$P\left[S\left(p_i^j\right)\right] = \begin{cases} 0, & if \ \ S\left(p_i^j\right) < \tau \\ 1, & otherwise \end{cases} \tag{3}$$

Among them, τ is used to measure the pixel brightness of the spatio-temporal subtitles, and pixels with a brightness value lower than τ will be regarded as interference and removed.

From Eq. (2), it can be known that the SSPA curve of the video can be formulated as the following Eq. (4).

$$SSPA = SSPA(1) \text{U } SSPA(2) \text{UL} SSPA(i) LU SSPA(L) \tag{4}$$

The algorithm for the building of SSPA curve is shown in Algorithm 1. The schematic diagram of SSPA curve is shown in Fig. 5.

(2) Catastrophe-Point Detection

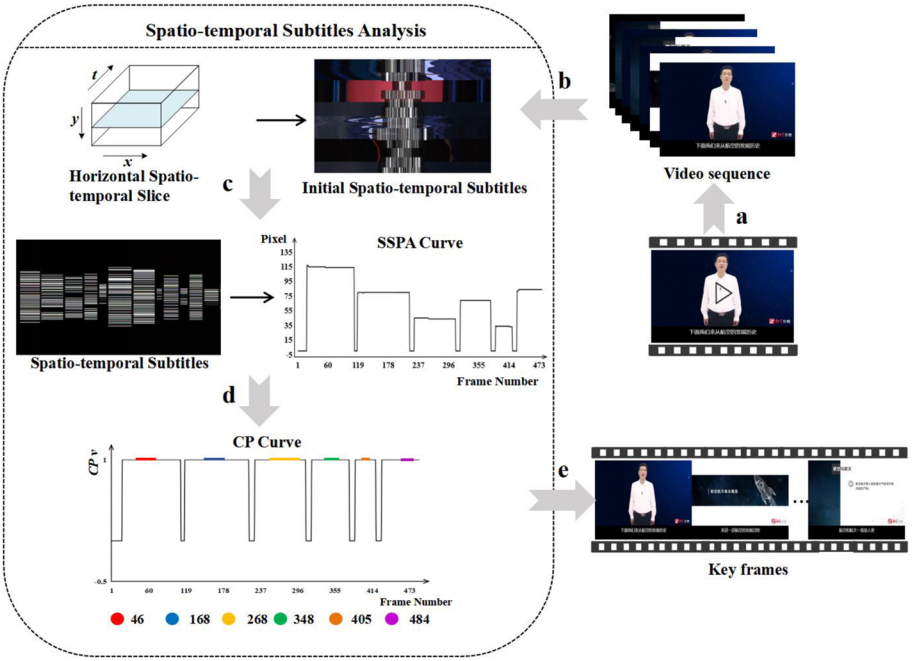

Fig. 4 Basic architecture of key frame extraction method based on spatio-temporal subtitles analysis. **a** Divide the input video into a series of single frame images according to the video frame rate. **b** Build the initial video spatio-temporal subtitle by using horizontal spatio-temporal slice techniques to extract the specific subtitle line pixels of a video image. **c** Establish SSPA curve by performing SSPA pixel statistics using the binarized processing maps of the initial spatio-temporal subtitle at a fixed threshold. **d** Extract the Catastrophe-Points of the SSPA curve in step c, and detect them by rising and falling edges, using the edge detection method of the cure. **e** Extract steady-state video key frames based on Catastrophe-Points and subtitle presence threshold

There are time gaps between subtitles, so the appearance of a new subtitle can cause a sudden change in the SSPA curve. Therefore, by detecting the Catastrophe-Point of the SSPA curve, the moment of appearance or disappearance of the frame with subtitle can be obtained. For simplicity, Catastrophe-Points are presented through a mapping of Catastrophe-Point values, where Catastrophe-Point values (denoted as *CP_v*)

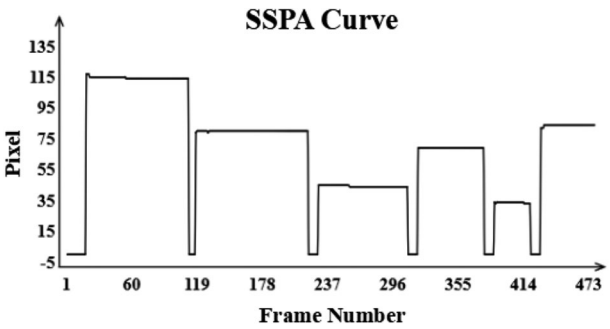

Fig. 5 The schematic diagram of SSPA curve

Input: Video $V(x,y,t)$, L %the length of the video $V(x,y,t)$

Output: SSPA curve

while $V(x,y,t) \neq \phi$ do

  Binarize(V) % the threshold is *mean*(pixel_ave$_{5\%}$)

  if $S(p_i^j) < \tau$ % the threshold of $S(p_i^j)$ is τ

    $P[S(p_i^j)] \leftarrow 0$

  else $P[S(p_i^j)] \leftarrow 1$

  end if

  $SSPA(i) = \sum_{j=1}^{W} P[S(p_i^j)]$ % the i^{th} frame in video $V(x,y,t)$

  $SSPA = SSPA(1) \quad SSPA(2) \quad \cdots SSPA(i) \cdots \quad SSPA(L)$

end while

function ***Binarize***(V)

  $LengthFiles \leftarrow L$

  for $i \leftarrow 1$ to $LengthFiles$ do

    $a \leftarrow read(V, i)$ % the following functions are all included in the platform

    $h \leftarrow sort(a(:), descend)$ % sort pixel values in descending order

    $v \leftarrow mean(h(1 : round(length(h) * 0.5)))$

    $Img \leftarrow a * (255 / v)$

  end for

  return Img

end function

Algorithm 1 Building SSPA curve

are calculated as shown in Eq. (5). When the CP_v is greater than the value 1, the Catastrophe-Point is considered to exist.

$$CP\_v = \begin{cases} 1 & SSPA(L) > w_0 \\ \frac{|SSPA(i+1) - SSPA(i)|}{w_0} & |SSPA(i+1) - SSPA(i)| > w_0 \\ 0 & otherwise \end{cases} \tag{5}$$

Among them, w_0 represents the threshold value of the significant difference between the SSPA value of the current subtitle frame and the previous subtitle frame. In order to select the key frame as comprehensively as possible, the proposed method defines the last frame of the video and the moment when the cumulative value of the pixel between the video frames is greater than the threshold w_0 as the Catastrophe-Point of the curve, and treats the CP_v at this moment uniformly according to Eq. (5), recording the CP_v in the rest of the cases as 0.

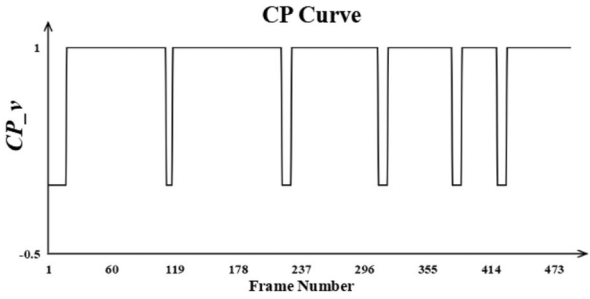

Fig. 6 The schematic diagram of CP curve

In Sect. 1, we mentioned that this paper designed a method of steady-state key frame selection, that is, select the subtitle frame with *CP_v* as 1 interval segment as the key frame, as shown in Fig. 6.

4 Experimental results and analysis

As discussed in Sect. 3, the proposed algorithm in this paper can quickly extract the steady-state key frames. In order to properly evaluate the performance of the proposed method in this paper, two aspects of the experiments were executed to verify its feasibility and superiority over the other methods. The experiments were performed on a general-purposed computer with Intel(R) Core(TM) i5-6200U processor with the main frequency of 2.30 GHz and the operating system is 64-bit Windows 10 Professional Edition, the algorithm experiment platform is Matlab2016b.

To ensure the generality of the method, 8 publicly available standard test videos of different scenes such as lecture, education, movie, sport and new were used in the experiments. These videos have embedded subtitles that facilitated identifying the difference among the tested key frame extraction methods. The detailed information of the video data set is shown in Table 1.

Section 4.1 verifies the feasibility of the proposed method by the test video example named video_T. Section 4.2 compares the objective performance of the proposed method with existing methods on 8 publicly available standard test videos and discusses the superiority of the proposed method compared with other methods.

Table 1 Specific information of the video data set

No	Video Name	Video Types	Number of Video Frame	Number of Key Frame
1	Bill Gates	Lecture	1000	12
2	The Journey of the Machine	Education	1434	19
3	Artificial Intelligence	Lecture	1826	26
4	Computer Network technology	Education	1205	16
5	Full River Red	Movie	896	8
6	Boil it up! Narrator	Sport	1102	13
7	Topics in Focus	New	1425	11
8	World Express	New	1228	18

4.1 Algorithm feasibility

In order to verify the feasibility of the proposed method, we selected a video clip, denoted as video_T, from an open course of Shanghai Jiao Tong University for analysis. The details of the video_T are as follows, with a duration of 15.9 s, a frame rate of 30 frames per second, and a total of 479 frames. According to the key frames defined in Sect. 2, the key frames of video_T containing the subtitle information are 6 frames, which are manually observed and counted.

In addition, the CP curve is an ascending and descending alternating curve obtained by curve edge detection, reflecting the presence and disappearance of subtitles. In the video, the same subtitle can get the same SSPA value in the ideal state, so the curve tends to stabilize in the corresponding interval, and there are two different states, namely, the falling edge of the subtitle from existence to disappear (i. e., the *CP_v* from 1 to 0), and the rising edge of the subtitle from disappearing to existence (i. e., the *CP_v* from 0 to 1).

Based on the above observation and analysis, there is obvious Catastrophe-Point in the SSPA curve of the video, which can be used as a basis for detecting subtitle changes. As can be seen from Fig. 7 that the SSPA curve of the video_T produces a momentary abrupt change at frames 19th, 112th, 117th, 220th, 229th, 310th, 379th, 388th, 421th, 430th of the video, that is, the Catastrophe-Points of the curve are located at these positions. Analysis and calculation conclude that the steady-stable key frames extracted based

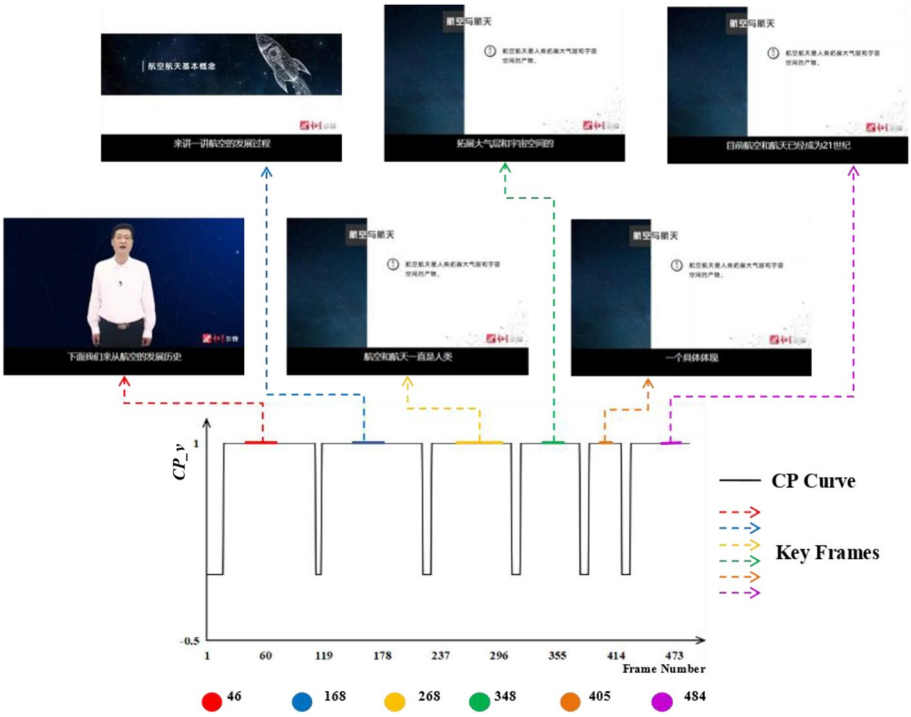

Fig. 7 Experimental results for the video_T

on the method proposed in this paper are the 46th, 168th, 268th, and 348th, 405th, 464th frames of the video, a total of 6 frames, which further confirms the feasibility of the algorithm in this paper.

4.2 Algorithm validity

The possible errors in key frame extraction are mainly manifested in the following three aspects: the first one is the missing selection,which the key frames extracted in the manual observation process are not extracted by the algorithm; the second one is the redundant selection, that is, the key frames containing key information can be represented by only a single frame, but the algorithm may select the same two or more frames, and the last one is the wrong selection, that means the key frames extracted by the algorithm do not contain the key information in the video.

For experiments on the key frame extraction of video, the number of missing selections, redundant and wrong selections of the key frame extraction are usually used to demonstrate the experimental results. In addition, this paper also uses four metrics, namely, precision, recall, F_1-score and running time, to evaluate and analyze the performance of the algorithm. The precision (Eq. (6)), recall (Eq. (7)), and F_1-score (Eq. (8)) were used to evaluated the performance of the proposed method.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F_1 - \text{score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

Among them, the true positive (TP) represents the number of frames correctly detected from all key frames, false positive (FP) represents the number of frames incorrectly detected, when it is not a key frame, and false negative (FN) represents the number of frames missing detected, when it is a key frame.

In order to analyze the validity of the proposed method, the experiments are compared with the classic method based on K-means Clustering [28], the method based on the adjacent frame difference [24] and the method based on the multimodal features semantic similarity [7]. The experimental results are shown in Table 2.

It can be seen from Table 2 that compared with other key frame extraction methods, the F_1-score composite indexes on the 8 test videos is 1, 0.97, 0.98, 1, 0.94, 1, 0.95 and 0.92, respectively. The average F_1-score composite index can reach 0.97, among which the average precision can reach 0.97 and the average recall can reach 0.98. Therefore, the algorithm in this paper can extract the key frames in steady-state and can represent the main contents of the video more comprehensively.

In terms of running time, the fastest running time of this algorithm is 9.12 s under video 4, which is significantly reduced compared with the running time of other algorithms. The algorithm extracts the key frames by analyzing the spatio-temporal subtitle features of the video, that is, using the spatio-temporal slice technique to analyze the single line pixels of the video frame, which successfully reduces the computational amount and speeds up the running speed compared with other algorithms that analyze the global features of the video frames.

Table 2 Comparative experimental results of the tested methods

Video	Algorithms	Key Frame	TP	FN	FP	P	R	F1-score	Time(s)
1	Method [28]	169	12	0	0	1	1	1	36.29
	Method [24]	69	9	3	0	1	0.75	0.86	1071.3
	Method [7]	15	8	2	0	1	0.8	0.89	598.9
	Method of this paper	12	12	0	0	1	1	1	36.01
2	Method [28]	236	18	1	31	0.367	0.947	0.53	24.78
	Method [24]	50	19	0	7	0.731	1	0.84	160.93
	Method [7]	18	12	5	3	0.8	0.71	0.37	765.8
	Method of this paper	19	18	0	1	0.97	1	0.97	11.24
3	Method [28]	150	10	16	17	0.37	0.385	0.38	13.53
	Method [24]	138	26	0	30	0.464	1	0.63	171.97
	Method [7]	30	21	0	8	0.724	1	0.42	811.6
	Method of this paper	26	25	0	1	0.962	1	0.98	13.45
4	Method [28]	229	16	0	40	0.286	1	0.44	17.93
	Method [24]	21	16	0	0	1	1	1	94.54
	Method [7]	29	14	5	5	0.737	0.737	0.37	766
	Method of this paper	16	16	0	0	1	1	1	9.12
5	Method [28]	126	8	0	21	0.276	1	0.432	33.58
	Method [24]	13	8	0	3	0.727	0.625	0.842	72.9
	Method [7]	18	5	3	6	0.455	1	0.526	329.4
	Method of this paper	8	8	0	1	0.889	1	0.941	21.6
6	Method [28]	96	12	0	15	0.445	1	0.615	23.5
	Method [24]	23	12	0	1	0.923	1	0.96	344.9
	Method [7]	35	9	3	9	0.5	0.75	0.6	406.3
	Method of this paper	12	12	0	0	1	1	1	19.4
7	Method [28]	114	9	0	23	0.281	1	0.9	37.7
	Method [24]	53	7	2	3	0.7	0.778	0.737	252.3
	Method [7]	65	6	5	3	0.667	0.545	0.6	362.2
	Method of this paper	9	10	1	0	1	0.91	0.952	35.7
8	Method [28]	187	15	2	19	0.441	0.882	0.9	29.8
	Method [24]	68	14	3	8	0.636	0.824	0.718	95.3
	Method [7]	79	17	6	11	0.607	0.739	0.67	702.5
	Method of this paper	17	17	2	1	0.944	0.895	0.92	26.1

In summary, the algorithm in this paper not only can effectively extract the key frames of the video and represent the video content more completely, but also achieves very high operation efficiency, and realize the key frames extraction of the lecture video, quickly and accurately.

5 Conclusion

In this paper, we propose a key frame extraction method for lecture videos based on spatio-temporal subtitle. The method uses spatio-temporal subtitle to analyze the video subtitles changes and selects the steady-state subtitle frames as key frames. With the SSPA curve,

the appearance and disappearance moments of video subtitles can be captured, and the key frames in steady-state are extracted based on curve edge detection and subtitle presence threshold. The test results of 8 videos demonstrate that the method in this paper outperforms existing methods with considerable effectiveness and robustness. Future work is to extend the algorithm of this paper to build adaptive spatio-temporal subtitles.

Acknowledgements This work is jointly supported by the National Natural Science Foundation of China (No.61702347, No.62027801); the Natural Science Foundation of Hebei Province (No.F2022210007, No.F2017210161); the Science and Technology Project of Hebei Education Department (No.ZD2022100, No.QN2017132); the Central Guidance on Local Science and Technology Development Fund (No.226Z0501G); the Shijiazhuang Tiedao University Graduate Innovation Funding Project under Grant (YC2022058).

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflicts of interest The authors declare no conflicts of interest to report regarding the present study.

References

1. Alhakami W, Binmahfoudh A, Baz A et al (2021) Atrocious impinging of covid-19 pandemic on software development industries. Comput Syst Sci Eng 36(2):323–338
2. Alrumiah SS, Al-Shargabi AA (2022) Educational videos subtitles' summarization using latent dirichlet allocation and length enhancement. Comput Mater Contin 70(3):6205–6221
3. Bai HR, Lü JL (2017) Improved algorithm of key frame extraction based on clustering methods. Comput Eng Des 38(07):1929–1933. https://doi.org/10.16208/j.issn1000-7024.2017.07.041
4. Bhardwaj AK, Garg L, Garg A, Gajpal Y (2021) E-learning during covid-19 outbreak: cloud computing adoption in Indian public universities. Comput Mater Contin 66(3):2471–2492
5. Chen DJ (2018) Investigating video-based listening test: the role of visual information in academic lecture comprehension process. Dissertation, Zhejiang University, China
6. Chen JY, Lao SY, Wu LD (2003) Video Abstraction. J Image Graph 4(07):3–7
7. Cui XD, Liu DW, Liu YF, Zhao ZB, Ren YG, Yan YM (2022) Research and Implementation of Key Frame Summarization Model for News Short Video. Comput Eng:1–9 https://doi.org/10.19678/j.issn.1000-3428.0065727
8. Hu ZJ, Xu Y (2020) Overview of Content-based Video Retrieval. Comput Sci 47(01):117–123
9. Hua R, Wu XX, Zhao WT (2021) Video summarization by learning semantic information. J Beijing Univ Aeronaut Astronaut 47(03):650–657
10. Jin K, Feng HC, Yang T (2014) Multi-modality Video Scene Segmentation Algorithm with Semantic Concept. J Chin Comput Syst 35(09):2156–2161
11. Jin H, Zhou YH (2000) Review of Video Parsing Techniques for Contented Based Video Retrieval. J Image Graph 5(04):276–283
12. Kethsy PA, Shree JD (2019) Histogram difference with Fuzzy rule base modeling for gradual shot boundary detection in video cloud applications. Clust Comput 22(1):1211–1218
13. Li YY, Wang JL (2020) Self-Attention Based Video Summarization. J Comput Aided Des Comput Graph 32(04):652–659
14. Liu YH (2014) Survey of Algorithms for Partitioning Video into Shots in Video. Technol Innov Appl 16:49–50
15. Lo CC, Wang SJ (2001) Video segmentation using a histogram-based fuzzy c-means clustering algorithm. Comput Stand Interfaces 23(5):429–438
16. Ma J, Yao A, Mao RZ (2021) Innovation in the Cultivation Model of Digital Literacy for Citizens in the Post-epidemic Era. Libr Inf 41(2):75–83
17. Pan XF, Li JT, Zhang YD, Tang S, Yu LJ, Xia T (2009) Spatiotemporal Video Copy Detection Based on Visual Perception Analyses. Chin J Comput 32(01):107–114
18. Qu Z, Gao TF, Zhang QQ (2012) Study on an Improved Algorithm of Video Keyframe Extraction. Comput Sci 39(08):300–303

19. Shan LY, Li XW (2019) Video Fingerprinting Algorithm Based on Temporal and Spatial Information Feature Fusion. Comput Eng 45(08):260–265+274. https://doi.org/10.19678/j.issn.1000-3428.0052107
20. Srinivasu PN, JayaLakshmi G, Jhaveri RH, Praveen SP (2022) Ambient assistive living for monitoring the physical activity of diabetic adults through body area networks. Mob Inf Syst 2022:1–18. https://doi.org/10.1155/2022/3169927
21. Su XH (2020) Research on key frame extraction and video retrieval from the perspective of deep learning. Netw Secur Technol Appl 233(05):65–66
22. Vulli A, Srinivasu PN, Sashank MSK, Shafi J, Choi J, Ijaz MF (2022) Fine-Tuned DenseNet-169 for Breast Cancer Metastasis Prediction Using FastAI and 1-Cycle Policy. Sensors (Basel) 22(8):2988. https://doi.org/10.3390/s22082988
23. Wang Y, Chen SJ (2014) News video anchor shot detection based on template and color moment. Wirel Internet Technol 06:161
24. Wang ZH, Li JT, Xie SY, Zhou J, Li HJ, Fan X (2018) Two-stage Method for Video Caption Detection and Extraction. Comput Sci 45(08):50–53+62
25. Wang HX, Wang L, Yan SS (2019) Research on Key Frame Extraction Method in Video Retrieval. J Shenyang Ligong Univ 38(03):78–82
26. Wu LS (2015) Research on Soccer Video Index Structures and Retrieval Algorithms. Dissertation, Huazhong University of Science and Technology, China
27. Xiang LY, Wang L, Hu HY, Mao H, Wang YM, Cha JL et al (2020) Quality control of online TCM teaching and learning during the fight against COVID-19. Educ Chin Med 39(03):27–31
28. Yin Y, Jiang HN (2007) Key frame extraction based on clustering of optimizing initial centers. Comput Eng Appl 43(21):165–167
29. Zhang QQ (2018) Surveillance Key-Frame Extraction with Spatio-Temporal Graph Representation. Dissertation, Anhui University, China
30. Zhang HJ, Wu JH, Di Z, Smoliar SW (1997) An integrated system for content-based video retrieval and browsing. Pattern Recogn 30(4):643–658
31. Zhang YZ, Zhang SS, Li Y, Zhang JY, Cai ZQ, Shui L (2021) Surveillance video key frame extraction based on center offset. Comput Mater Contin 68(3):4175–4190
32. Zhang XR, Zhang WF, Sun W, Sun XM, Jha SK (2022) A Robust 3-D Medical Watermarking Based on Wavelet Transform for Data Protection. Comput Syst Sci Eng 41(3):1043–1056
33. Zhong MJ, Zhang YB (2019) A Key Frame Extraction Method of Vehicle Surveillance Video Based on Visual Saliency. Comput Technol Dev 29(06):164–169