Computers, Materials & Continua
DOI:10.32604/cmc.2021.017011
Article

Tech Science Press

Surveillance Video Key Frame Extraction Based on Center Offset

Yunzuo Zhang[1,*], Shasha Zhang[1], Yi Li[1], Jiayu Zhang[1], Zhaoquan Cai[2] and Shui Lam[3]

[1]School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang, 050043, China
[2]Department of Computer Science and Engineering, Huizhou University, Huizhou, 516007, China
[3]College of Engineering, California State University, Long Beach, CA, 90185, USA
[*]Corresponding Author: Yunzuo Zhang. Email: zhangyunzuo888@sina.com
Received: 18 January 2021; Accepted: 24 March 2021

Abstract: With the explosive growth of surveillance video data, browsing videos quickly and effectively has become an urgent problem. Video key frame extraction has received widespread attention as an effective solution. However, accurately capturing the local motion state changes of moving objects in the video is still challenging in key frame extraction. The target center offset can reflect the change of its motion state. This observation proposed a novel key frame extraction method based on moving objects center offset in this paper. The proposed method utilizes the center offset to obtain the global and local motion state information of moving objects, and meanwhile, selects the video frame where the center offset curve changes suddenly as the key frame. Such processing effectively overcomes the inaccuracy of traditional key frame extraction methods. Initially, extracting the center point of each frame. Subsequently, calculating the center point offset of each frame and forming the center offset curve by connecting the center offset of each frame. Finally, extracting candidate key frames and optimizing them to generate final key frames. The experimental results demonstrate that the proposed method outperforms contrast methods to capturing the local motion state changes of moving objects.

Keywords: Center offset; local motion; key frame extraction; moving object detection

1 Introduction

With the gradual improvement of people's security awareness, the demand for a surveillance video system is increasing sharply in recent years. The rapid development of video surveillance system has brought about the explosive growth of video data. Seeking an object in such lengthy videos is similar to looking for a needle in a haystack. Such a dilemma requires efficient video managing and browsing to solve [1–4]. Key frame extraction enables users to effectively browse and manage massive videos, which aroused considerable interest.

Key frame extraction aims to remove redundant frames and remain as important video information as [4] to achieve efficient video management and facilitate users to quickly query

and browse videos. Currently, existing commonly key frame extraction algorithms include: shot boundary-based, content change-based, clustering-based, motion analysis-based and so on. Typical key frame extraction methods based on shot boundaries [5–7] select the first or the last frame of each shot as the key frame. The method is easy to implement. Nevertheless, it may result in unacceptable results when video motion intensity is excessive since the first frame and the last frame do not necessarily represent the key content of the video shot. Content-based analysis methods extract key frames according to visual information such as color, texture, and shape. For instance, Zhang et al. [8] utilized the color histogram difference between the current frame and the next key frame to select key frames. Gunsel et al. [9] used the color histogram difference between the current frame and previous N frames to extract key frames. These methods are effective. However, they may not extract potential key frames well when there are many changes in the video. Besides that, regardless of how effective the low-level feature-based methods are, the loss of detailed semantics is virtually inevitable. The reason is that these methods extract the key frames based on a single frame or a small number of frames. There are also some methods to extract key frames based on clustering, and these methods choose the nearest frame from each cluster center as the key frames. In the clustering-based methods, Zhuang et al. [10] clustered frames according to color histogram similarity and extracted the key frames from the clusters. Hanjalic et al. [11] proposed a method for generating key frames and previews abstract forms for an arbitrary video sequence. The underlying principle of the proposed method is to eliminate the visual-content redundancy among video frames. The method accomplishes this by applying multiple partitional clustering to all frames of a video sequence and then selecting the most suitable clustering options using an unsupervised cluster-validity analysis procedure. In the last step, select key frames as the centroids of obtained optimal clusters. The clustering-based methods can generate an acceptable video summary, but in contrast, the method usually requires a high computation cost. Furthermore, the method will cause the sequential order of the key frames may not be preserved. In addition to the above work, many deep learning methods [12–14] are also used to extract key frames. For example, RPCA-KFE is presented in [14], a key frame extraction algorithm that considers both the contribution to video reconstruction and the distinctness of each video frame. The disadvantage of these methods is due to their massive calculation.

Among the widely used key frame extraction methods, motion-related ones have demonstrated good performances, meanwhile, these will be discussed in detail in Section 2. Nevertheless, these methods only focus on global object motion state changes such as starting, stopping, accelerating, decelerating, or direction changing. When applied in scenes with local motion such as bending over and stretching up, they are not as good as expected.

The changes in the local motion state of moving objects can arouse more attention, especially in surveillance videos. Local motion can be accurately reflected by the center offset of moving objects, thus we define the frames with the maximum local center offset of moving objects as key frames, and propose a key frame extraction method. To the best of our knowledge, there is no such published work considering this issue. Therefore, it is interesting and worthwhile to research.

The remainder of this paper is arranged as follows. Section 2 briefly reviews several previous motion-related key frame extraction methods and motion object detection methods. Section 3 explains the concept of objects center offset and describes the framework of the proposed method. Experimental results of the proposed method and contrast methods on various video sequences are presented in Section 4. Finally, conclusions are provided in Section 5.

2 Related Work

2.1 Motion-Related Key Frame Extraction Method

In this section, we review the traditional motion-related key frame extraction methods. As an effective method to solve the problem of large video data browsing, key frame extraction has been widely used in surveillance video applications. A comprehensive and detailed investigation of the existing key frame extraction methods has been made in [15,16].

There are numerous key frame extraction methods based on motion analysis. Wolf [17] first calculated the optical flow for each frame to set a motion metric and then analyzed the metric as a function of time to select the key frames. This method can select key frames appropriate to the composition of the video shot. However, considerable computation is required to calculate the optical flow. Liu et al. [18] put forward the hypothesis that motion is a more salient feature in presenting actions or events in videos. Based on this hypothesis, a triangle model based on perceived motion energy (PME) represents the motion activities in video shots. Liu et al. [19] addressed key frame extraction from the viewpoint of shot reconstruction degree (SRD) and proposed an inflexions-based algorithm. This algorithm first calculates each frame's motion energy to form a curve and then uses polygon simplification to search the inflexions of the energy curve; finally, the frames which at the inflexions of the energy curve are extracted as key frames. It shows effective performance in fidelity and SRD; however, the inflexions of the energy curve are not the same as the inflexions of the video sequence. Ma et al. [20] proposed a new key frame extraction method based on motion acceleration. This method uses motion acceleration of the primary moving object to obtain the motion state changes, such as start, stop, acceleration, deceleration, or direction change. The key frames extracted by this method can describe the changes of the motion state. Li et al. [21] presented a motion-focusing method to extract key frames, which focuses on one constant-speed motion and aligns the video frames by fixing this focused motion into a static situation. According to the relative motion theory, the other video objects are moving relative to the selected kind of motion. Zhong et al. [22] proposed a fully automatic and computationally efficient framework for analysis and summarization of surveillance videos. This framework uses the motion trajectory to represent the moving process of the target. Zhang et al. [23] presented a method for key frame extraction based on spatio-temporal motion trajectory, which can obtain the state changes of all moving objects. This method defines frames at inflexions of motion trajectory on the spatiotemporal slice (MTSS) as key frames. The reason is that the inflexions of the MTSS can capture all motion state changes of moving objects.

The above methods can all show excellent performance under the circumstances, however, they tend to ignore the changes in the local motion state of the moving objects. The center offset of moving objects can be employed to describe the changes of the local motion state. Under this observation, the paper proposed a key frame extraction method based on center offset.

2.2 Moving Object Detection

As one of the most fundamental and challenging problems in object extraction, object classification [24–26], object tracking [27], crowd counting [28] and object recognition [29], objection detection has attracted considerable attention in recent years. Many papers on moving object detection have been published. A study on various methods used for moving object detection in video surveillance applications has been made in [30].

As a hot topic in video processing, moving object detection plays a vital role in the subsequent processing of object classification, tracking, and behavior understanding in videos. However, due

to the complex video scenarios, there are still many problems with moving object detection needed to be solved. Currently, the background subtraction method and frame difference method are two common methods for moving object detection in surveillance videos. The background subtraction method's basic steps [31] are as follows: firstly, establishing the background model and then comparing the input image with it. Finally, moving objects are detected by the statistical information changes such as gray level or histogram.

The conventional inter-frame difference method is to subtract two consecutive adjacent frames to obtain moving objects. If a pixel is very different from the surroundings, it is usually caused by moving objects in the video frame. If these pixels are marked, the moving objects in the video frames can be obtained. This method is simple, and the amount of the calculation is not very large, but the obtained moving object maybe with "holes". Therefore, some scholars have improved the traditional inter-frame difference method, and the more effective one is the three-frame difference method.

In addition to the above methods, there are optical flow methods [32,33], background modeling method, etc. Combining the advantage of various moving object detection algorithms [34] can reach a good detecting result. Based on the analysis of the above methods and the experimental videos' actual scene, this paper adopts the background difference method with background updating to detect moving objects.

3 Proposed Method

3.1 Center Offset

The center point of each moving object shape is defined as the center point, which can also be called the centroid in mathematics. Mathematically, the centroid of a curved surface is the geometric center of the cross-section figure, and the centroid is the centroid of the abstract geometry. For objects with uniform density, the center of mass coincides with the centroid. In the process of motion, the moving object in the video may have different action behaviors. We think that the moving object is an abstract geometry with uniform density, changing its shape constantly. Therefore, different cross-section shapes will be left in each video frame during the moving process. The cross-section shape generated in each frame, called the object motion shape, as shown in Fig. 1.

Fig. 1 shows the shape formed by the target when the target is doing erect, reaching, squatting, etc. From Fig. 1, we find that when a moving object makes local motion, the motion shape will change, that is, the position of the object center will be offset. That is why the center offset is employed to reflect the changes of the local motion state.

Next, how do we calculate the center coordinates of the object moving shape? In the Cartesian coordinate system, if the coordinates of the vertices of the triangle are (x_1, y_1), (x_2, y_2), and (x_3, y_3), respectively, the coordinates of midpoint (x, y) can be calculated as:

$$x = \frac{1}{3}(x_1 + x_2 + x_3), \quad y = \frac{1}{3}(y_1 + y_2 + y_3) \tag{1}$$

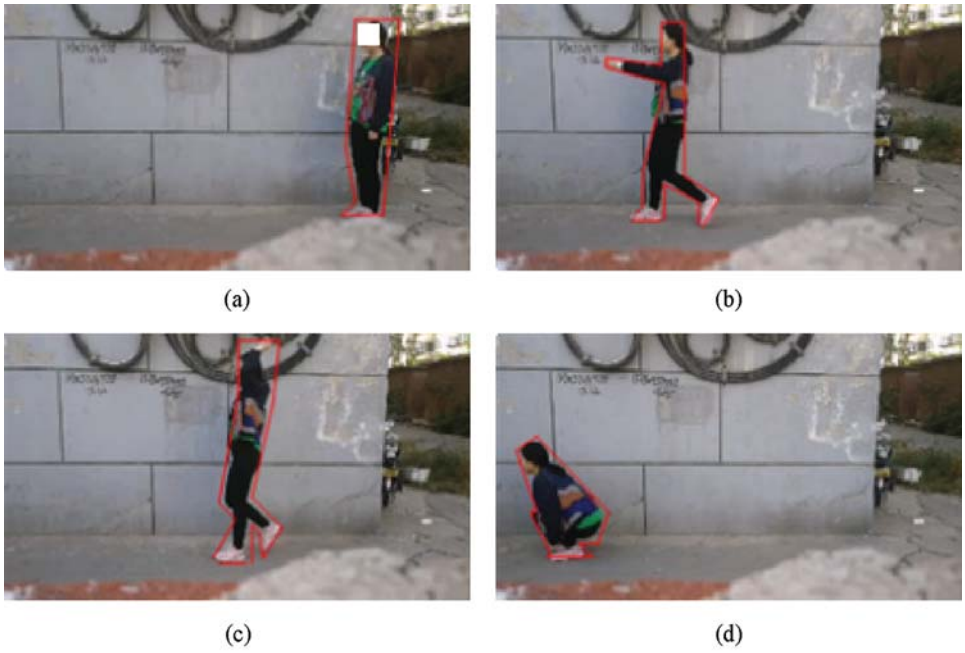

Figure 1: Moving shape of the object (a) Erecting (b) Rasing hand (c) Walking (d) Squatting

If the fixed-point coordinates of the rectangle are (x_1, y_1), (x_1, y_2), (x_2, y_1) and (x_2, y_2), respectively, then the coordinates of midpoint (x, y) can be obtained by:

$$x = \frac{1}{2}(x_1 + x_2), \quad y = \frac{1}{2}(y_1 + y_2) \tag{2}$$

When the figure is a polygon, the double integral is needed to calculate the centroid. To simplify the calculation, the center point of the circumscribed rectangle of moving object is selected to represent moving object. An example of the center point is shown in Fig. 2.

Figure 2: An example of the center point

From Fig. 2, it can be seen that the position of rectangle center point changes with the position of the rectangle's four vertices. It indicates that when the moving object makes local motion such as bending over or stretching up, it will cause the outer rectangle changes, and the position of outer rectangle center point will change accordingly. That is, the object center offset can reflect the changes of both global and local motion state of the moving object. Therefore, we select the center offset of outer rectangle as the motion descriptors, and use it to describe the changes of all motion states.

When there is only one moving object in the video frame, the center point of the object rectangle is the center point of the video frame. However, when there are multiple moving objects in the video frame, the center point of video frame is the average value of the center points of each object, as shown in Fig. 3.

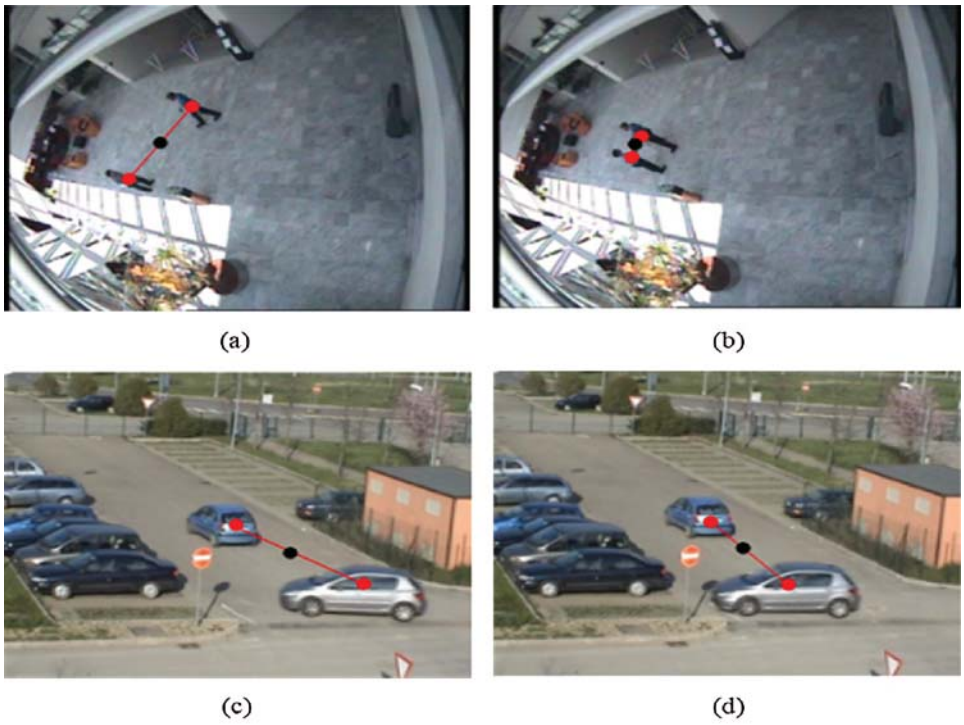

Figure 3: Center point of multi-object video frame (a) Initial center point of pedestrian (b) Center point after moving of pedestrian (c) Initial center point of vehicles (d) Center point after moving of vehicles

From Fig. 3, it can be found that when multiple objects are moving at the same time, the average value of each object rectangle center point is used as the center point of the video frame. The reason is that when one object moves, its center point will change, the coordinates of the frame center point will change too, so the center offset of the video frame can reflect the changes of each object's motion state. Therefore, it is feasible to use the center offset of moving objects in adjacent video frames to reflect the changes of moving state of objects.Under this observation, a video key frame extraction method based on moving target center offset is proposed.

For video V, the center offset of moving object can be defined as:

$$\boldsymbol{CO}(t) = \boldsymbol{CO}_x(t) + \boldsymbol{CO}_y(t) \quad \boldsymbol{CO}(t) = \boldsymbol{CO}_x(t) + \boldsymbol{CO}_y(t), \tag{3}$$

where $\boldsymbol{CO}(t)$ represents moving object center point offset at time t, $\boldsymbol{CO}_x(t)$ and $\boldsymbol{CO}_y(t)$ are the horizontal component and the vertical component of $\boldsymbol{CO}(t)$, respectively. Let $P(x_1, y_1, t-1)$ and $P(x_2, y_2, t)$ denote the coordinates of moving object center point at times $t-1$ and t, respectively. Then the center offset $\boldsymbol{CO}(t)$ can be expressed as:

$$\boldsymbol{CO}(t) = P(x_2, y_2, t) - P(x_1, y_1, t-1) \tag{4}$$

The vector in Eq. (4) can be computed as:

$$\boldsymbol{CO}(t) = |\boldsymbol{CO}(t)| \ \exp[-j\theta(t)] \tag{5}$$

where $|\boldsymbol{CO}(t)|$ and $\theta(t)$ represent the magnitude and angle of the $\boldsymbol{CO}(t)$, respectively. Where:

$$|\boldsymbol{CO}(t)| = \sqrt{(x_2 - x_1) + (y_2 - y_1)^2} \tag{6}$$

Eq. (5) shows that when $|\boldsymbol{CO}(t)|$ is large enough, it is easy to be extracted as a key frame. However, it does not only depend on $|\boldsymbol{CO}(t)|$, but also $\exp[-j\theta(t)]$ is a very important factor. For simplicity, $\exp[-j\theta(t)]$ is defined as:

$$\exp[-j\theta(t)] = \begin{cases} 8 & \text{moving direction reverses,} \\ 4 & \text{start or stop,} \\ 1 & \text{else.} \end{cases}$$

The center offset of each video frame can be calculated by using the above equation.

3.2 Key Frame Extraction Based on Center Offset

This paper defines the frame where the center shift peak abruptly changes as a key frame. Accordingly, a novel key frame extraction method based on center offset is proposed. The framework of the proposed method is shown in Fig. 4.

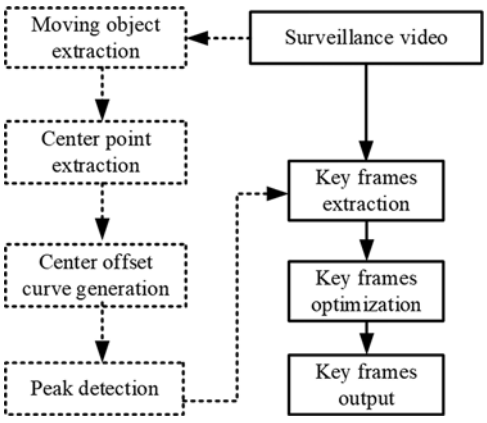

Figure 4: The framework of proposed method

Step 1. Moving object extraction

Firstly, it uses the background subtraction method to detect the moving object in the input surveillance video sequence, then extracts the moving object, and finally marks the moving object with the circumscribed rectangle.

Step 2. Center point extraction

It selects the midpoint of the circumscribed rectangle of moving object as the object center to obtain the coordinate value of the object center.

Step 3. Center offset curve generation

It calculates the center offset of the object by using the center point coordinates which have been known, and then connects the center offset of each frame to form a center offset curve.

Step 4. Peak detection

The peak of the center offset curve formed in Step 3 is detected, and the video frame corresponding to the peak value of the curve are extracted as candidate key frames.

Step 5. Key frames extraction

In order to reduce the redundancy of key frames, it needs to extracted the video frame where the peak value of the current frame is N times that of the previous key frame. Finally, the extracted video frames are composed of the video frames at the peak mutation, the first frame and the last frame of the input surveillance video.

Next, optimize the extracted key frames according to the visual resolution mechanism [35] to determine the final key frames.

In practice, the key frame number k will be extracted to ensure the objectivity. When the number of extracted key frames K (i.e., the final number of key frames determined in (Step 5)) is less than the specified number of key frames, the video frames with larger peak value except for the key frames are inserted first. If the video frames at other peak points are not enough for K–K frames, the remaining video frames are used to make up for the missing ones by interpolation method [36]. On the contrary, the smaller peak (K–K) frames in the final key frames determined in Step 5 are removed, and the specified key frames number K is extracted.

4 Experimental Results and Analysis

To correctly evaluate the correctness and effectiveness of the proposed method, we executed the experiments to verify its validity and superiority over the state of the art methods. The experiments were performed on a general-purpose computer with an Intel Core (TM) i5-4200 CPU and 8 GB memory.

4.1 Experiment Preparation

4.1.1 Test Data Set

The experiment used 16 test videos of different scenes to ensure the generality of the method. Some of them are from standard data set ViSOR [37], CAVIAR [38], and BEHAVE [39], and others are self-collected surveillance videos. Tab. 1 shows the detailed information of the above test video.

Table 1: Information on test videos

Video	Name	Fps	Number
Video 1	Prova	25	109
Video 2	Video 2	25	135
Video 3	Visor_1212744706330_type3_p1_small	25	212
Video 4	Visor_1268244090985_car1_0	7	54
Video 5	Visor_1212674142673_pacco1	25	92
Video 6	Fight_Chase	25	437
Video 7	LeftBag	25	1446
Video 8	Meet_WalkTogether 1	25	714
Video 9	Fight_OneManDown	25	965
Video 10	59800-66750	25	7410
Video 11	Visor_1292838684828_CWSv7	25	215
Video 12	Visor_1205423649326_Video00	25	2755
Video 13	Visor_1246523233130_new_8_camera 1	10	410
Video 14	OneLeaveShopReenter2cor	25	560
Video 15	Cam3_120405A	25	961
Video 16	Squating	30	36

4.1.2 Evaluation Criterion

To demonstrate the correctness and effectiveness of the proposed method, subjective and objective evaluation criteria are all used in this experiment. Subjective criteria mainly include result discussion and user studies, and the widely used objective evaluation criteria are Fidelity [40] and SRD [19]. Compared with the Fidelity criterion, the SRD criterion can evaluate the key frames from the dynamic aspect of capturing local details. If it has high SRD, it must have high Fidelity. Nevertheless, high fidelity does not necessarily mean high SRD. Therefore, the result discussion can verify the correctness of the proposed method, and comparative analysis and SRD criteria to verify the effectiveness.

4.2 Correctness

To demonstrate the correctness of the method, we applied the proposed method to 16 test videos and achieved desirable results. To be specific, the extracted key frames indicated that frames with the global and local motion state of objects, in a variety of scenes, could be effectively extracted by the proposed method. Due to space limitations, the article only takes the two key frame extraction results in Figs. 5 and 6 (corresponding to Video 8 and Video 16 in Tab. 1) as examples to illustrate the correctness. These examples are representative in terms of scenes and objects.

Fig. 5 displays the extracted result of video 8. Video 8 indicates two men walking face to face in a hall, shaking hands, and then walking together. Set the experimental parameter to $N = 5$.

This method discards some video frames with high peaks and extracts video frames with relatively low peaks as key frames. This is the result of the parameter setting and optimization criteria. In detail, due to the influence of environmental changes, the video frames before No. 58 have a higher peak value. Therefore, we optimize the experimental results by setting parameters and key frame optimization. The frames after No. 148 got lower peak values due to the parameter

settings. By setting the parameters, we have extracted some key frames. This ensures that extracted key frames can describe the whole motion of video 8. In this video, the scenes that attracted more attention were the appearance of two targets, the handshake of the two targets, and the changes in their movement after the handshake. Observing the result of key frame extraction, we can found that frame No. 58, No. 117 and No. 136 respectively show the appearance of two objects (the changes of global motion state of the objects), and frames No. 148, No. 172, No. 187 show the process of reaching out before handshake (the change of local motion state of objects). Frames No. 206 and No. 229 show the handshake of two targets (the change of local motion state of objects, and frame No. 251 shows two objects moving in another direction after handshake (the change of global motion state of objects).

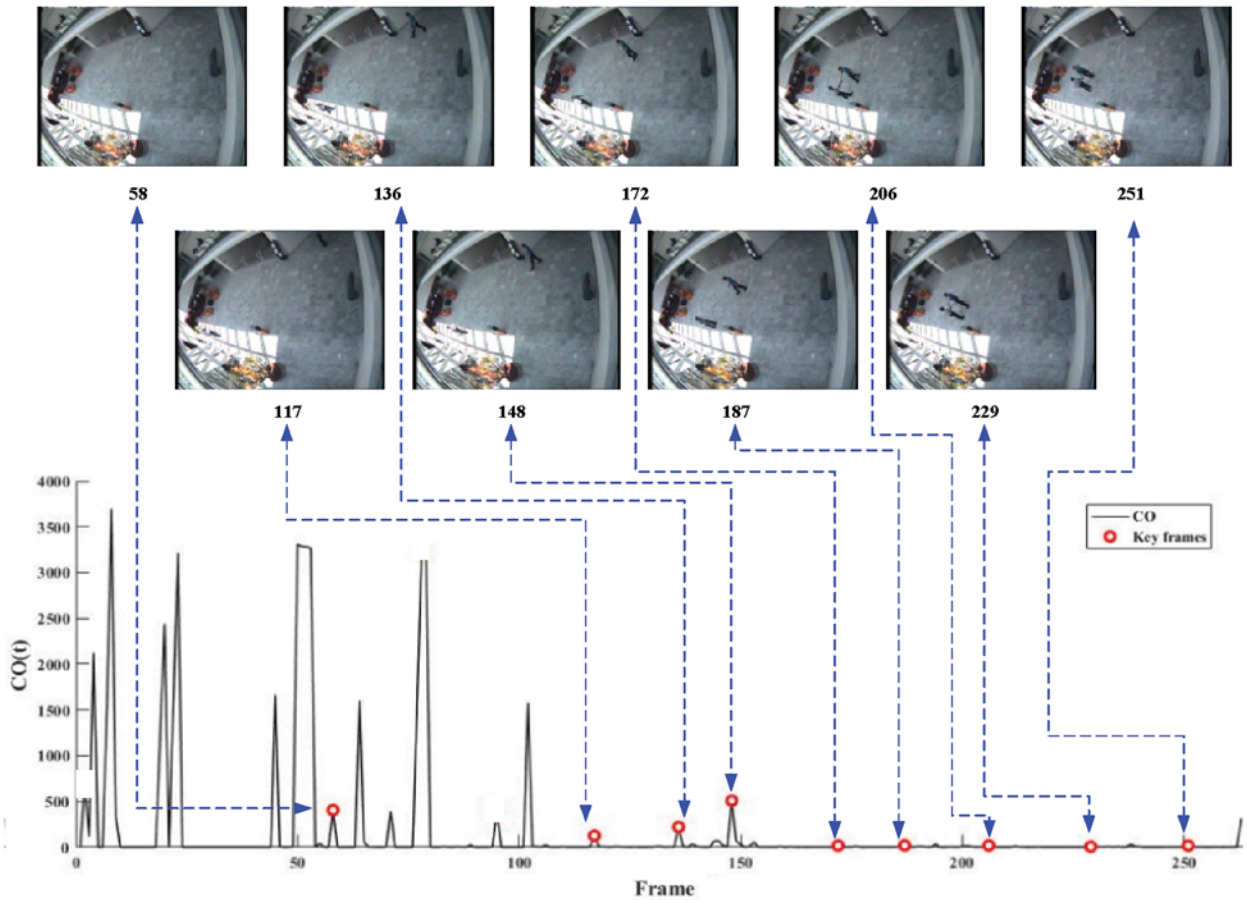

Figure 5: Video key frames extracted from video 8

Fig. 6 presents the extracted key frames result of video 16 (the final key frames except for the first frame and the last frame). Video 16 indicates that a woman does the movement of standing-squatting-standing up.

Fig. 6 shows the similar results. The video frames at the first peak and the last peak are calculated according to the optimization criteria, which are similar to the first frame and the last frame. The proposed method can extract the target squatting action (frame No. 6), the target

squatting action change (frame No. 13) and the target standing up action (frames No. 21, No. 26, No. 29). This demonstrates that the proposed method can extract the changes of the local motion state well.

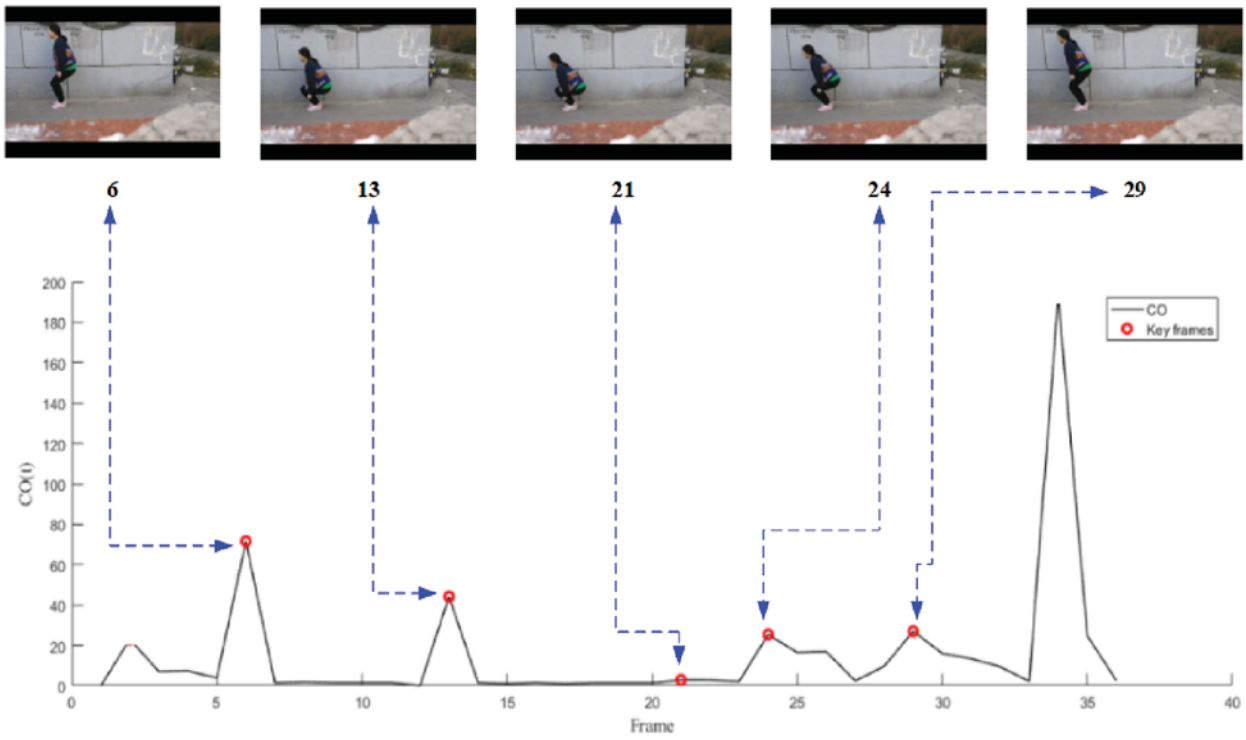

Figure 6: Video key frames extracted from video 16

As discussed above, we extract the video key frames based on the attractive feature of local and global motion state changes, and obtain them by analyzing the offset of the target center. Consequently, they are consistent with human visual perception. The discussion in this subsection validates the correctness of the proposed method.

4.3 Effectiveness

To verify the effectiveness of the proposed method, also test it with state of the art motion-related methods. The experiment compares the method proposed in this article with other methods such as the method based on the perceptual motion energy model in [18] (denoted as ME), the method based on motion acceleration (denoted as MA) in [20], and the method based on spatiotemporal motion trajectory In [23], it is expressed as MTSS for comparison, and in [41], the method based on motion speed (denoted as MV) is implemented. They are closely related to the proposed method. To ensure the universality and robustness of the proposed method, experiments were conducted on 16 test videos of the public data set and self-collected video. The performance comparison of the proposed method with the other methods was using the subjective criterion and objective criterion. The details are presented as follows.

The key frame extraction results of the five methods are firstly evaluated using the subjective criterion. In order to ensure the objectivity of the experiment, every method extracts 10 frames

as key frames. Through the test on 16 video segments, the results of five key frame extraction methods are obtained. The proposed method is superior to the others. Due to the limited number of pages, only the key frame extraction results of video 8 are displayed. The key frame extraction results of the proposed method and contrast methods of video 8 sequence are shown in Fig. 7.

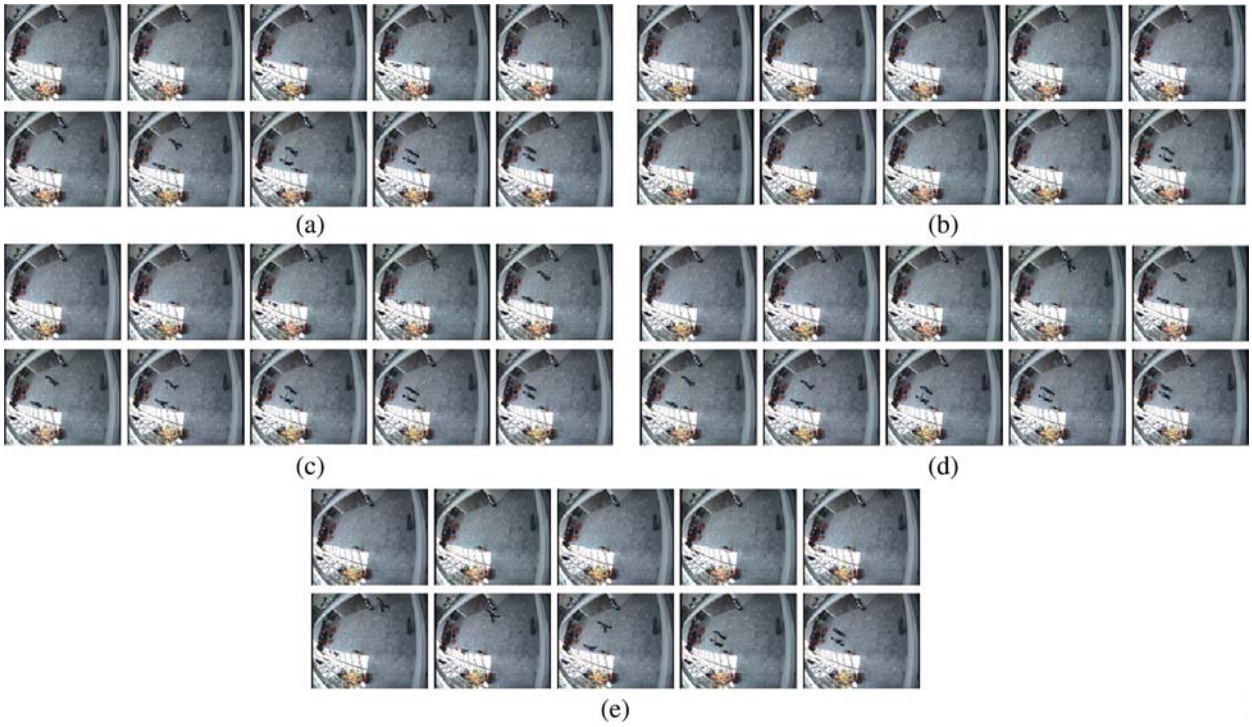

Figure 7: Key frame extraction results of video 8 (a) Proposed (b) MV (c) MA (d) MTSS (e) ME

From Fig. 7, it can be seen that the proposed method extracted the process of two targets appearing separately and shaking hands. MV can extract the video frames of the two targets and the handshake action of the two targets, but the video frame extracted by this method omits the movement process before the handshake. The key frames extracted by MTSS omitted the video frame in which the first moving object appears. The key frames extracted by MA omitted the video frame of the first moving object appearing, and it failed to extract the video frame of the second moving object appearing. The key frames extracted by ME have much redundancy and blank frames. To sum up, the proposed method can extract the video frames with motion state changes in it, especially in the multi-object surveillance video.

As an objective criterion, SRD is used to evaluate the key frame extraction results of the proposed method and its contrast methods. SRD criterion is to evaluate the key frame extraction method from the aspect of video reconstruction ability. The larger the calculated SRD is, the better the video reconstructs. This means that the video reconstructed by the extracted key frames is closer to the original video. Fig. 8 shows the average SRD obtained by the proposed method and its contrast methods on all test videos with different key frame ratios (2% –12%).

From Fig. 8, it can be seen that the average SRD increases with the number of key frames which are extracted by the proposed method and contrast methods. When the key frame extraction

rate is 2% to 6%, the average SRD of the proposed method is almost the same as that of MV, MTSS and MA, and it is significantly higher than that of ME. When the key frame extraction rate is between 8% and 12%, the average SRD of the proposed method is about 0.3dB higher than contrast methods. The reason is that the proposed method considers the local and global changes of all object motion states, while contrast methods only focus on the global motion state changes. It can be concluded that the proposed method is superior to contrast methods in SRD criterion, and the proposed method can capture the local motion state changes of each object better. Therefore, the above discussion demonstrates that the proposed method is effective.

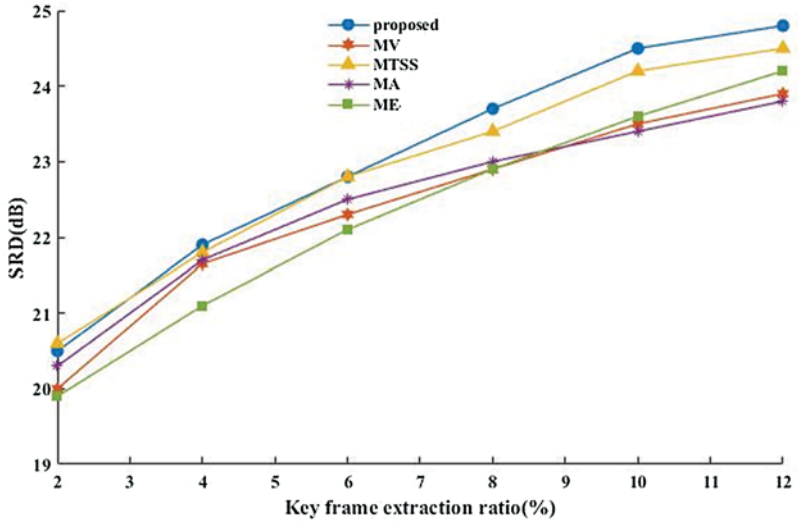

Figure 8: The average SRD of the proposed method and its contrast methods

5 Conclusions

This paper proposed a novel center offset-based extraction method to extract the key frame in the surveillance video. The center offset is used to capture the global and local motion state changes of moving objects. In other words, it means to replace the object with the center point of the moving target. When there are multiple objects in the video frame, this method calculates the mean value of the center point of these moving targets as the center point of the video frame. Next, calculate the center offset of each frame and then connect them to form a center offset curve. Finally, extract the video frame at the peak mutation as the key frame. Experimental results demonstrate that the proposed method outperforms the existing state-of-the-art methods in capturing the local motion state changes of moving objects.

Funding Statement: This work was supported by the National Nature Science Foundation of China (Grant No. 61702347, 61772225), Natural Science Foundation of Hebei Province (Grant No. F2017210161).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Edison and C. V. Jiji, "Optical acceleration for motion description in videos," in *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, Honolulu, USA, pp. 1642–1650, 2017.

[2] H. Gharbi, S. Bahroun and E. Zagrouba, "Key frame extraction for video summarization using local description and repeatability graph clustering," *Signal Image and Video Processing*, vol. 13, no. 4, pp. 507–515, 2019.

[3] K. Kim, "An efficient implementation of key frame extraction and sharing in android for wireless video sensor network," *KSII Transactions on Internet and Information Systems*, vol. 9, no. 9, pp. 3357–3376, 2015.

[4] Y. Luo, H. Zhou, Q. Tan, X. Chen and M. Yun, "Key frame extraction of surveillance video based on moving object detection and image similarity," *Pattern Recognition and Image Analysis*, vol. 28, no. 2, pp. 225–231, 2018.

[5] A. Nagasaka and Y. Tanaka, "Automatic video indexing and full-video search for object appearances," in *IFIP Tc2/wg 26 Second Working Conf. on Visual Database Systems II*, Budapest, Hungary, pp. 113–127, 1991.

[6] Y. Rui, T. S. Huang and S. Mehrotra, "Exploring video structure beyond the shots," in *Proc. IEEE Int. Conf. on Multimedia Computing and Systems*, Austin, TX, USA, pp. 237–240, 1998.

[7] J. H. Yuan, H. Y. Wang, L. Xiao, W. J. Zheng, J. M. Li *et al.,* "A formal study of shot boundary detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 2, pp. 168–186, 2018.

[8] H. J. Zhang, J. Wu, D. Zhong and S. W. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognition*, vol. 30, no. 4, pp. 643–658, 1997.

[9] B. Gunsel and A. M. Tekalp, "Content-based video abstraction," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, IL, USA, vol. 3, pp. 128–132, 1998.

[10] Y. Zhuang, Y. Rui, T. S. Huang and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. Int. Conf. on Image Processing*, Chicago, IL, USA, vol. 1, pp. 866–870, 1998.

[11] A. Hanjalic and H. J. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1280–1289, 1999.

[12] S. H. Mei, G. L. Guan, Z. Y. Wang, S. Wan, M. Y. He *et al.,* "Video summarization via minimum sparse reconstruction," *Pattern Recognition*, vol. 48, no. 2, pp. 522–533, 2015.

[13] K. Zhang, W. L. Chao, F. Sha and K. Grauman, "Summary transfer: Exemplar-based subset selection for video summarization," in *Proc. Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1059–1067, 2016.

[14] C. Dang and H. Radha, "RPCA-KFE: Key frame extraction for video using robust principal component analysis," *IEEE Trans. Image Process*, vol. 24, no. 11, pp. 3742–3753, 2015.

[15] W. M. Hu, N. H. Xie, L. Li, X. L. Zeng and M. Stephen, "A survey on visual content-based video indexing and retrieval," *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 6, pp. 797–819, 2011.

[16] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 3, no. 1, pp. 1–37, 2007.

[17] W. Wolf, "Key frame selection by motion analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal*, Atlanta, USA, vol. 2, pp. 1228–1231, 1996.

[18] T. M. Liu, H. J. Zhang and F. H. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 10, pp. 1006–1013, 2003.

[19] T. Y. Liu, X. D. Zhang, J. Feng and K. T. Lo, "Shot reconstruction degree: A novel criterion for key frame selection," *Pattern Recognit. Lett.*, vol. 25, no. 12, pp. 1451–1457, 2004.

[20] Y. Z. Ma, Y. L. Chang and H. Yuan, "Key-frame extraction based on motion acceleration," *Opt. Eng.*, vol. 47, no. 9, pp. 957–966, 2008.

[21] C. C. Li, Y. T. Wu, S. S. Yu and T. Chen, "Motion-focusing key frame extraction and video summarization for lane surveillance system," in *Proc. the 16th IEEE Int. Conf. on Image Processing*, Cairo, Egypt, pp. 4329–4332, 2009.

[22] J. Zhong, Y. Su, R. Qian and J. Ma, "Surveillance video summarization based on moving object detection and trajectory extraction," in *Proc. Int. Conf. on Signal Processing System*, Dalian, China, pp. 250–253, 2010.

[23] Y. Z. Zhang, R. Tao and F. Zhang, "Key frame extraction based on spatiotemporal motion trajectory," *Opt. Eng.*, vol. 54, no. 5, pp. 1–3, 2015.

[24] C. L. Du, S. H. Liu, L. Si, Y. H. Guo and T. Jin, "Using object detection network for malware detection and identification in network traffic packets," *Computers, Materials & Continua*, vol. 64, no. 3, pp. 1785–1796, 2020.

[25] A. Qayyum, I. Ahmad, M. Iftikhar and M. Mazher, "Object detection and fuzzy-based classification using UAV data," *Intelligent Automation & Soft Computing*, vol. 26, no. 4, pp. 693–702, 2020.

[26] J. C. Chen, Z. L. Zhou, Z. Q. Pan and C. N. Yang, "Instance retrieval using region of interest based CNN features," *Journal of New Media*, vol. 1, no. 2, pp. 87–99, 2019.

[27] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang *et al.,* "T-CNN: Tubelets with convolutional neural networks for object detection from videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2896–2907, 2018.

[28] Z. Q. Xiao, B. Yang and D. Tjahjadi, "An efficient crossing-line crowd counting algorithm with two-stage detection," *Computers, Materials & Continua*, vol. 60, no. 3, pp. 1141–1154, 2019.

[29] H. Yoon, K. J. Kim and J. Chun, "Shadow detection and removal from photo-realistic synthetic urban image using deep learning," *Computers, Materials & Continua*, vol. 62, no. 1, pp. 459–472, 2020.

[30] A. S. Murugan, K. S. Devi, A. Sivaranjani and P. Srinivasan, "A study on various methods used for video summarization and moving object detection for video surveillance applications," *Multimedia Tools and Applications*, vol. 77, no. 18, pp. 23273–23290, 2018.

[31] S. Y. Chiu, C. C. Chiu and S. D. Xu, "A background subtraction algorithm in complex environments based on category entropy analysis," *Appl. Sci.*, vol. 8, no. 6, pp. 885, 2018.

[32] H. Sidenbladh, "Detecting human motion with support vector machines," in *Proc. the 17th Int. Conf. on Pattern Recognition*, Cambridge, UK, vol. 2, pp. 188–191, 2004.

[33] K. Kuzume and M. Okada, "Sensor network system to promote energy conservation realization of energy smart school," in *Proc. IEEE Int. Conf. on Pervasive Computing and Communication Workshops*, Budapest, Hungary, pp. 187–190, 2014.

[34] X. W. Han, Y. Gao, Z. Lu, Z. M. Zhang and D. Niu, "Research on moving object detection algorithm based on improved three frame difference method and optical flow," in *Proc. 2015 Fifth Int. Conf. on Instrumentation and Measurement, Computer, Communication and Control*, Qinhuangdao, pp. 580–584, 2015.

[35] Y. Zhang, R. Tao and Y. Wang, "Motion-state-adaptive video summarization via spatio-temporal analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 6, pp. 1340–1352, 2017.

[36] M. Shio, M. Yanagisawa and N. Togawa, "Linear and bi-linear interpolation circuits using selector logics and their evaluations," in *Proc. IEEE ISCAS*, Melbourne VIC, Australia, pp. 1436–1439, 2014.

[37] R. Vezzanii and R. Cucchiara, "Video surveillance online repository (ViSOR): An integrated framework, Kluwer Academic Publishers, [Online]. Available: http://imagelab.ing.unimore.it/visor/.

[38] F. Robert, "CAVIAR Test Case Scenarios," 2011. [Online]. Available: http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/.

[39] S. J. Blunsden and R. B. Fisher, "The BEHAVE video dataset: Ground truthed video for multi-person behavior classification," *Annals of the BMVA*, no. 4, pp. 1–12, 2010. [Online]. Available: http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/.

[40] H. S. Chang, S. Sull and S. U. Lee, "Efficient video indexing scheme for content-based retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1269–1279, 1999.

[41] Y. Z. Zhang, S. S. Zhang and Y. N. Guo, "Research on the key-frame extraction of surveillance video based on motion velocity," in *Proc. 2019 IEEE Int. Conf. on Signal, Information and Data Processing*, Chongqing, China, pp. 1–3, 2019.