Journal of Information Hiding and Multimedia Signal Processing ©2019 ISSN 2073-4212
Ubiquitous International Volume **10**, Number **2**, March **2019**

A Fast Key Frame Extraction Method for The Surveillance Video Based on The Moving Targets

Jia-Qi Gui and Zhe-Ming Lu*

School of Aeronautics and Astronautics
Zhejiang University
Hangzhou, 310027, P. R. China
zheminglu@zju.edu.cn

Received October 2017; revised July 2018
(Communicated by Zhe-Ming Lu)

Abstract. *Considering the problems that the timeliness of the key frame extraction method using the comparison of the Cumulative Global Color Histogram*[1] *(CGCH) combined with the comparison of central moment of each original block is not high, our paper proposes a fast key frame extraction method based on the moving targets in surveillance video, which makes use of surveillance where a lot of redundancy is present. First of all, our paper uses the foreground extraction based on Gaussian mixture model (GMM)to obtain the foreground which contains the moving objects in surveillance video. Eroding and dilating and 4-neighborhood searching algorithm are performed to confirm the position of moving objects. From this, we can select the foreground frames and rearrange them to form a more brief video sequence. Finally, our paper adopts the algorithm using the comparison of the CGCH combined with the comparison of central moment of each original block to extract the key frames. Experimental results indicate that the proposed methods outperform the original method in both extraction accuracy and processing speed specially for surveillance where a lot of redundancy is present.*

Keywords: Cumulative Global Color Histogram, Central moment, GMM foreground extraction, 4-neighborhood searching algorithm, Key frame extraction.

1. **Introduction.** Video surveillance has been widely used in many applications. Public safety and theft protections are most important uses of it. A system like this needs an efficient transmission and storage of the large video data. Key frame extraction is a simple and powerful system to accomplish this objective[2]. Key frame extraction from video refers to extracting the frames that can represent the original video content according to predefined criteria, which can reduce the redundancy in video data and only retain the most useful parts of video data. The current key frame extraction method can be divided into the following three categories. (1) Key frame extraction based on shot boundary[3, 4]. The method firstly divides the video into several shots and takes the shot as the basic unit. Then the method selects the first frame and the last frame of the shot as the key frames and extracts a fixed number of frames in the shot as the key frames according to the judgment model. (2)Clustering-based key frame extraction[5]. The main idea of this method is to group frames with similar low-level features and select the frame closest to each cluster center as a key frame[6, 7, 8]. Though clustering is a good method for key frame extraction, it has some drawbacks such as it does not consider temporal information[9, 10]. (3)Key frame extraction based on motion. The motion is an intrinsic attribute of video and human eyes are very sensitive to it; thus they take into account motion events and camera

operations in key frame extraction. The triangular model based on Perceived Motion Energy (PME)[11] is a typical motion-based key frame extraction method. Since the method does not require threshold judgment, the calculation speed is very fast. Ma et al.[12] assume that the change of motion states attracts more attention than motion itself. They define the frames with the most significant acceleration (MSA) as key frames. However, surveillance video taken by the monitoring equipment has no obvious video structure and shot transition. Therefore, some traditional key frame extraction methods are not suitable to deal with the surveillance video.

According to the characteristics of the surveillance videomost of the video frames are the same background images (The content is not compact enough, the redundancy is high), our paper using the GMM foreground detection, eroding and dilating and 4-neighborhood searching algorithm extracts the surveillance video sequence which contains the moving objects(This method can reduce the length of the sequence from the video which processed by the key frame extraction algorithm). Finally, our paper adopts the algorithm using the comparison of the CGCH combined with the comparison of central moment of each original block to extract key frames in the simplified surveillance video sequence. The algorithm makes full use of global and local information in the frame. Compared with the algorithm using the comparison of the CGCH combined with the comparison of central moment of each original block, experimental results indicate that this algorithm deals with the surveillance video more quickly, and its extracted key frames express the contents of the surveillance video more briefly.

2. The Moving Target Detection in The Surveillance Video.

2.1. The Algorithm Using GMM Foreground Detection.

Background modeling using mixture of Gaussians is a classical algorithm of basic background subtraction[13]. The GMM algorithm for foreground detection is mainly used to initialize the surveillance video and extract the foreground areas in all the frames. In this technique, each pixel of a scene is modeled independently by a mixture of at most K Gaussian distributions. With the arrival of the new images, parameters (average, mean and weight) of Gaussian distributions are continually updated[14], and each pixel has to be matched to Gaussian distributions to determine whether it is updated or not. So it can accurately characterize the background information in real time. Compared with the single Gaussian model, it can commendably deal with dynamic background which is regularly changing as well as individual and mutational background models.

The formula of the Background modeling using mixture of Gaussians is as follows:

$$P(x_t) = \sum_{i=1}^{K} \omega_{i,t} \times \eta_{i,t}(x_t, \mu_{i,t}, \Sigma_{i,t}) \tag{1}$$

Where x_t is the sample value of the pixel at time t; $K \in \{3,4,5\}$is the number of Gaussian distributions in the model; $\omega_{i,t}$ is the weight parameter of the i^{th} Gaussian distribution at time t,and the sum of K weights is 1; $\mu_{i,t}$ is the mean of the i^{th} Gaussian distribution at time t; $\Sigma_{i,t}$ is the covariance of the i^{th} Gaussian distribution at time t; $\eta_{i,t}$ is a Gaussian probability density function defined as:

$$\eta_{i,t}(x_t, \mu_{i,t}, \Sigma_{i,t}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_{i,t}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_t - u_{i,t})^T \Sigma^{-1} (x_t - \mu_{i,t})} \tag{2}$$

Where n is the dimension of x_t.Based on the independent assumptions of color, covariance is defined as $\Sigma_{i,t} = \sigma_{i,t}^2 I$(here, $\sigma_{i,t}^2$ is the standard deviation of i^{th} Gaussian distribution at time t). If one pixel value in a frame is satisfied with $|x_t - \mu_{i,t-1}| < D \times \sigma_{i,t-1}$, in other words, the current pixel x_t matches the i_{th} Gaussian, so $M_{i,t} = 1$ and when unmatched, $M_{i,t} = 0$. D is a constant threshold identical to 2.5[15], which controls the rigorous level of foreground extraction. The smaller the value, the more stringent the demands are. The parameters are updated by the following formulas.

$$\omega_{i,t} = (1-\alpha)\omega_{i,t-1} + \alpha(M_{i,t}) \tag{3}$$

$$\mu_{i,t} = \begin{cases} (1-\rho)\mu_{i,t-1} + \rho x_t & when \quad M_{i,t} = 1 \\ \mu_{i,t-1} & when \quad M_{i,t} = 0 \end{cases} \tag{4}$$

$$\sigma^2_{i,t} = \begin{cases} (1-\rho)\sigma^2_{i,t-1} + \rho(x_t - \mu_{i,t})^T(x_t - \mu_{i,t}) & when \quad M_{i,t} = 1 \\ \sigma^2_{i,t-1} & when \quad M_{i,t} = 0 \end{cases} \tag{5}$$

$$\rho = \alpha\eta(x_t \mid \mu_{i,t}\ ,\ \sigma_{i,t}) \tag{6}$$

Where α is the learning rate ($0 \le \alpha \le 1$) and the value of α decides the background frame update rate; ρ is the learning rate of the $\sigma^2_{i,t}$ and $\mu_{i,t}$. If $M_{i,t} = 0$, $\sigma^2_{i,t}$ and $\mu_{i,t}$ remain unchanged and the pixel belongs to foreground, and a new Gaussian distribution is established to replace the original Gaussian distribution whose priority is the smallest. This method uses the value of the nearest pixel as the average of the new Gaussian distribution, then it initializes a smaller weight and a larger variance. With the time going by, for an updated mixture model, if one pixel has always matched one distribution of K Gaussian distributions, *i.e.*, $M_{i,t} = 1$. Then as the time goes on, ω will constantly increase and σ will constantly keep decreasing. By sorting for ω/σ, ω is normalized again. By setting weights and threshold, we cut out the former b Gaussian distributions with the highest weight, which are taken as the background model:

$$B = \arg\min_b\left(\sum_i^b \omega_{i,t} > T\right) \tag{7}$$

Where b is a parameter from 1 to K, T is a threshold value chosen high for multi-model distribution with repetitive motion in background. If the threshold T is small, the model is often a single Gaussian model, which is the best Gaussian distribution (the weight is the largest). If the threshold T is big, it will use multiple distributions as models, and it is stable for scenes such as shaking of leaves, lakes ripples, and so on. The pixel is judged to the background pixel as long as it matches any one of the former b Gaussian distributions. Otherwise, the pixel belongs to foreground.

2.2. **4-Neighborhood Searching Algorithm.** 4-neighbors, as the name implies, is the four neighbors of the pixelA pixel P at coordinates (x, y) has four horizontal and vertical neighbors whose coordinates are given by:(x+1,y), (x-1, y), (x, y+1), (x,y-1), as shown in the Fig.1.

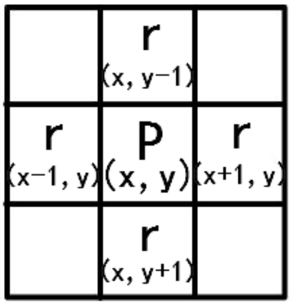
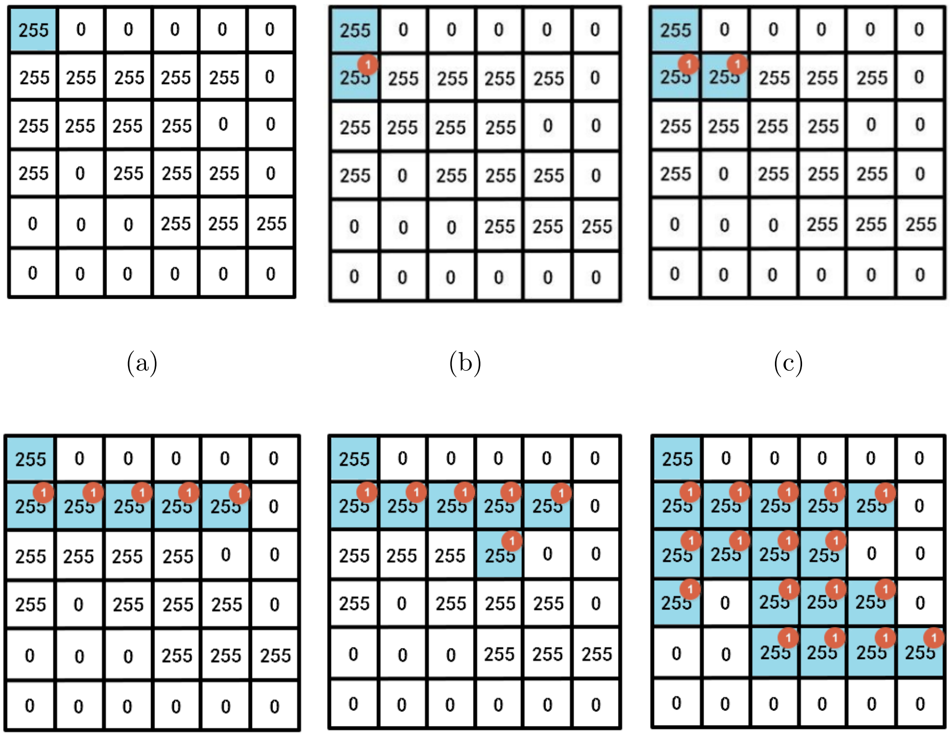


Figure 1. Four points are four neighbors of the center P



255	0	0	0	0	0
255	255	255	255	255	0
255	255	255	255	0	0
255	0	255	255	255	0
0	0	0	255	255	255
0	0	0	0	0	0

(a)

255	0	0	0	0	0
1 255	255	255	255	255	0
255	255	255	255	0	0
255	0	255	255	255	0
0	0	0	255	255	255
0	0	0	0	0	0

(b)

255	0	0	0	0	0
1 255	1 255	255	255	255	0
255	255	255	255	0	0
255	0	255	255	255	0
0	0	0	255	255	255
0	0	0	0	0	0

(c)

255	0	0	0	0	0
1 255	1 255	1 255	1 255	1 255	0
255	255	255	255	0	0
255	0	255	255	255	0
0	0	0	255	255	255
0	0	0	0	0	0

(d)

255	0	0	0	0	0
1 255	1 255	1 255	1 255	1 255	0
255	255	255	1 255	0	0
255	0	255	255	255	0
0	0	0	255	255	255
0	0	0	0	0	0

(e)

255	0	0	0	0	0
1 255	1 255	1 255	1 255	1 255	0
1 255	1 255	1 255	1 255	0	0
1 255	0	1 255	1 255	1 255	0
0	0	1 255	1 255	1 255	1 255
0	0	0	0	0	0

(f)

Figure 2. A 4-neighborhood searching example

The prerequisite for the 4-neighborhood searching method is that the image must be binary, *i.e.*, the value of foreground is 255 and the value of background is 0. The basic idea of method is to group all the adjacent foreground points to form the area. First, one or more pixels found in segmentation area are the starting point of the search, *i.e.*, initial seeds. Then the pixels in the 4-neighbors which have the same properties with the initial seeds are merged into one region. Moreover, we regard these pixels which merged into the region as new seed pixels to continue the above process until there is no pixel merging. In this way, a region has grown. The 4-neighborhood searching example is shown in Fig.2(a).The number in the grid indicates the value of pixels in the binary image, and the light blue pixel is the initial seed point. We assume that the initial seed point is visited in the order up, left, right, down, and next we find that the light blue point is the top-left corner without the left and the top pixels, so it is visited in the order right, down. Then the pixel below the initial seed point is the same as the initial seed point, therefore we merge them into a region and mark the point, which has been scanned, as shown in Fig.2(b).The next step is scanning the 4-neighbors in order up, left, right, down, and we get Fig.2(c) and Fig.2(d). In Fig.2(e), in addition to the scanned left point, the pixel values of the other three points are 0 in 4-neighbors. Hencewe have to go back to the last scan point and find its neighbors which have no mark and whose value is 255. We scan out the pixel which meets the requirements and get Fig.2(e). When all the pixels that the values are 255 have been marked, they(all the light blue blocks) are combined into the foreground area, as shown in Fig. 2(f).

So the foreground area is extracted by the 4-neighborhood searching, we can determine the maximum and minimum of its vertical and horizontal coordinates, *i.e.*, the boundary of the area. The 4-neighborhood searching method establishes the foundation for positioning the moving target area in the surveillance video.

3. KeyFrame Extraction Algorithm in Surveillance Video.

3.1. The Algorithm Based on The Comparison of Central Moment of Each Original Block.

(1) central moment: The central moment is proposed by Stricker and Orengo, whose key idea is that any color distribution in the image can be represented by its moments. The central moment is an effective and simple color feature, and most of the information of color distribution is concentrated on lower-order moments(Generally, only the first moment (mean), second (variance), third (skewness) are used to approximately estimate the overall color distribution of the image.). It does not require color quantization and its dimension is much lower than the color histogram. Whats more, experiment by Liu et al. proved that the effect of central moment is almost as big as the color histogram. Hence we use the central moment to perform the key frame extraction. The formula for the third moment is:

$$m_{3i} = (\frac{1}{N}\sum_{j=1}^{N}|p_{ij} - m_{1i}|^3)^{1/3} \tag{8}$$

Where m_{3i} is the pixel skewness value of the i^{th} sub-block; N is the number of pixels in the i^{th} sub-block; p_{ij} is the pixel value of the j^{th} bit in the i^{th} sub-block; m_{1i} is the pixel mean value of the i^{th} sub-block.

(2) blocking: When all shot frames are filtering, the purpose of blocking is to generate a redundant candidate key frame sequence. The result of blocking is to automatically select a different number of candidate key frames according to the amount of exercise of the moving target or the range of camera operation, such as scaling or panning. The candidate key frames can be used as a basis for judging the intensity of the movement of each shot.

(3) The comparison of differences: The formula of the comparison of the central moment difference is:

$$P(Q,D) = \sqrt{\omega_H\sum_{k=1}^{3}(m_{kh}^Q - m_{kh}^D)^2 + \omega_S\sum_{k=1}^{3}(m_{ks}^Q - m_{ks}^D)^2 + \omega_V\sum_{k=1}^{3}(m_{kv}^Q - m_{kv}^D)^2} \tag{9}$$

Where ω_H, ω_S, ω_V are the weighting coefficients. m_{kh}^Q: if $k = 1$, m_{1h}^Q is the pixel mean value of H channel of a sub-block in current frame. If $k = 2$, m_{2h}^Q is the pixel variance value of H channel of a sub-block in current frame. If $k = 3$, m_{3h}^Q is the pixel skewness value of H channel of a sub-block in current frame; m_{kh}^D: if $k = 1$, m_{1h}^D is the pixel mean value of H channel of a sub-block in the key frame. If $k = 2$, m_{2h}^D is the pixel variance value of H channel of a sub-block in the key frame.If $k = 3$, m_{3h}^D is the pixel skewness value of H channel of a sub-block in the key frame (The sub-blocks position of m_{kh}^D corresponds to the sub-blocks position of m_{kh}^Q); Similarly, m_{ks}^Q represents S channel and m_{kv}^Q represents V channel.

The method using the comparison of central moment of each original block: First, the first frame of the original video is used as the first frame of the candidate key frame sequence, and the subsequent frames use it as reference. The sub-blocks of the subsequent frames, in turn, are compared with the sub-blocks of the reference-frame by the formula of the comparison of differences. In order to reduce the amount of computation, two thresholds are set respectivelythe large threshold of the sub-blocks δ_1 and the small threshold of the sub-blocks $\delta_2(\delta_1 > \delta_2)$. When the difference between the sub-blocks of the subsequent frame and the sub-blocks of the reference-frame are greater than or equal to δ_1 or the number of sub-blocks whose difference is greater than δ_2) is to m, stopping the calculation. The result shows that the frame has a large change relative to the reference-frame. The subsequent frame is placed in the key frame candidate set, and it is regarded as a new reference-frame, and so on. When the comparisons of all the frames are finished, we get the candidate key frame sequence. Because the method using the comparison of central moment of each original block is particularly sensitive to local changes, so the redundancy of the candidate key frame sequence is higher. The subsequent work is to reduce the redundancy by the comparison of the CGCH to achieve the best results.

3.2. **The Algorithm Using The Comparison of The Cumulative Global Color Histogram.** The cumulative histogram based on the feature statistics of image is a one-dimensional discrete function, the formula is:

$$I(k) = \sum_{i=0}^{k} \frac{n_k}{N} (k = 0, 1, ..., L - 1) \tag{10}$$

Where k is the eigenvalue of the image; L is the range of the eigenvalue; n_k is the number of pixels with the eigenvalue k in the image, and N is the total number of image pixels. The cumulative histogram can greatly reduce the number of zero values that appear in the statistical histogram, so the distance between two colors on the characteristic axis will be proportional to their similar degree. In this paper, only the H component is used to make comparison, and the comparison formula chooses the Euclidean distance.

The method using the comparison of the CGCH: First, the first frame of the candidate key frame sequence is used as the first frame of the endpoint frame sequence, and the subsequent frames use it as reference. The subsequent candidate frame, in turn, is compared with the reference-frame by the formula of the Euclidean distance. When the distance between two frames is greater than or equal to , the subsequent frame is used as the second frame of the endpoint frame sequence and it is regarded as a new reference-frame, and so on. We repeat the above operations until the last frame of the candidate key frame (The last frame is directly used as the last frame of the endpoint frame sequence). Finally, we get the whole endpoint frame sequence.

4. **The Proposed Scheme.** Key frame extraction for surveillance video usually faces with the following major issues: 1. In order to retain the whole key frame sequence, the traditional key frame extraction method will introduce a large number of redundant frames; 2. Some algorithms with the high computational complexity, such as the clustering-based algorithm and Wolf's optical flow method, cannot meet real-time requirements. According to the characteristics of the surveillance videomost of video frames are the same background images and only a small part of video frames contain the moving objects, our paper proposes a new key frame extraction algorithm for surveillance video. First our algorithm uses the GMM foreground detection to do a simple operation for the surveillance videoExtracting the foreground image in the video. Then, the moving targets in the foreground image are positioned by using the eroding and dilating and 4-neighborhood searching algorithm. We extract a short surveillance video sequence which contains the moving objects through the positioning operation. The above procedures remove a large number of background frames, so we do a preliminary simplified operation for the surveillance video. Finally, the key frames in the simplified video sequence are extracted by the method using the comparison of the CGCH combined with the comparison of central moment of each original block. Our algorithm can be divided into the following stages as Fig.3:

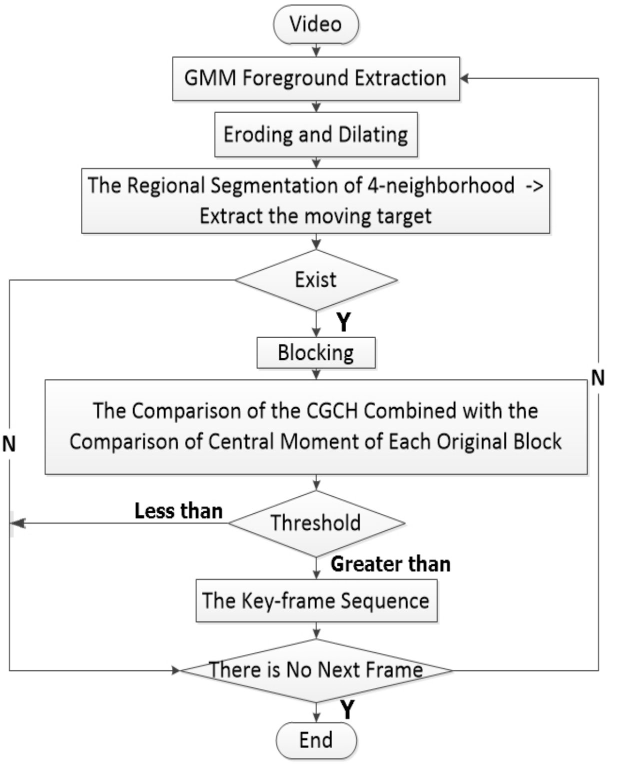
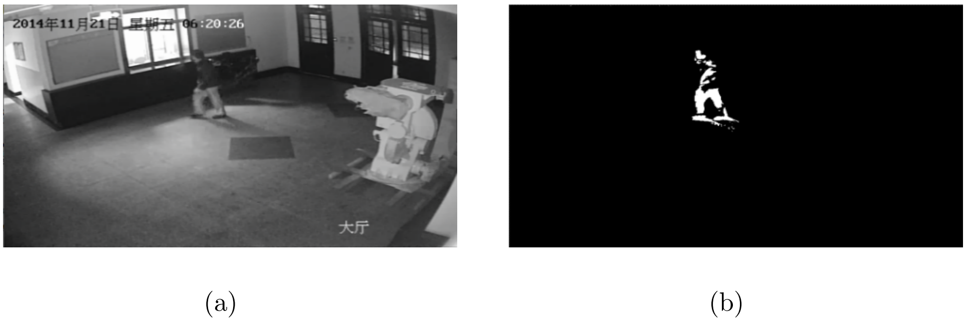

Video
GMM Foreground Extraction
Eroding and Dilating
The Regional Segmentation of 4-neighborhood -> Extract the moving target
Exist
Y
N
Blocking
The Comparison of the CGCH Combined with the Comparison of Central Moment of Each Original Block
N
Less than
Threshold
Greater than
The Key-frame Sequence
There is No Next Frame
Y
End

FIGURE 3. Flow chart of the algorithm framework



(a) (b)

FIGURE 4. Original image in the video and the image of the foreground extraction

Surveillance video itself is flooded with a lot of redundant information, only a small part of the key frames carry useful information and most of the video frames are the same background frames. In this paper, we first use the GMM foreground extraction algorithm to extract the foreground frames containing the moving objects, and reassemble them to form a new video sequence.

In the GMM algorithm, although it can achieve foreground detection, the effects of the foreground extraction will be poor, that is, the contour of object is sparse and there are many holes in the foreground, as shown in Fig.4(a) and Fig.4(b). By contrast, the algorithm adding the process of eroding and dilating can get the fuller object and suppress noise more effectively, as shown in Fig.5.

The method using eroding and dilating combined with the 4-neighborhood search algorithm can effectively detect the moving regions in the surveillance video, and the different

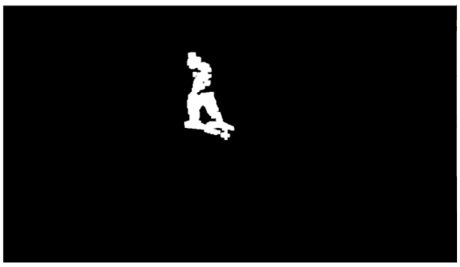


Figure 5. The algorithm with eroding and dilating

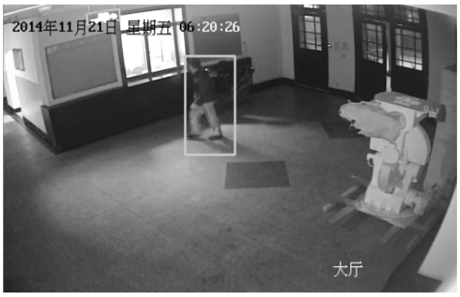


Figure 6. The result of the boundary extraction using 4-neighborhood search algorithm

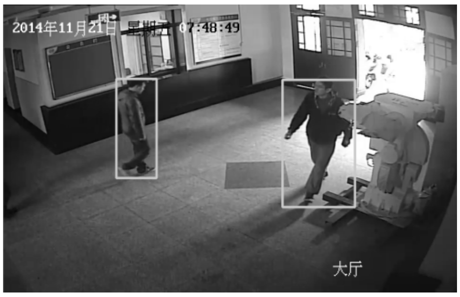


Figure 7. The result of the boundary extraction using 4-neighborhood search algorithm

moving regions are separated, as shown in Fig.6 and Fig.7. The number of moving objects, i.e., the number of borders, will determine whether or not to regard this frame as an effective frame in surveillance video.

According to the number of moving objects in the foreground, we can delete the redundant background frames and get a new video sequence. And then we process the each frame in the new video with blocking, as shown in Fig.8.

Our algorithm divides the video frame into 16(4×4) small blocks and calculates the first, second and third central moment of 16 sub-blocks in order. First, we regard the first frame of the original video as a reference frame, *i.e.*, it is used as the first frame of the candidate key frame sequence. The 16 sub-blocks of each subsequent frame, in turn, are compared with the corresponding sub-blocks of the reference-frame by the formula of the comparison of the central moment difference. One frame has 16 sub-blocks, so it has 16 differences after the calculation. If one of the differences is greater than or equal to δ_1(The large threshold) or the number of sub-blocks whose difference is greater than δ_2(The small threshold) is to m, we put the corresponding subsequent frame into the key

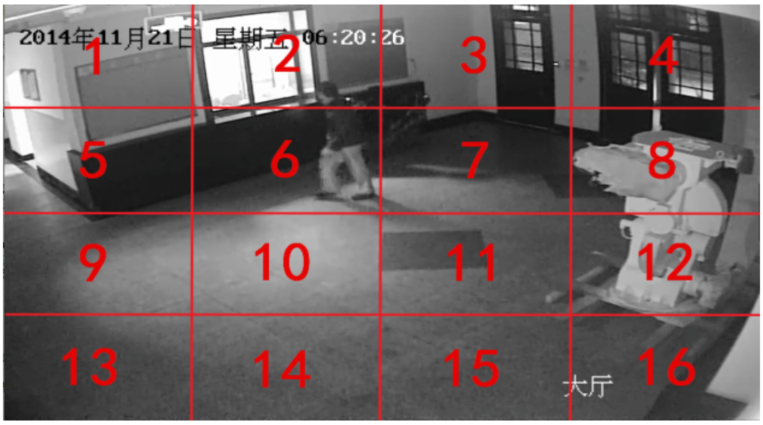


Figure 8. The blocking effect for the frame in surveillance video

frame candidate set and regard it as a new reference frame, and so on. When all the frames in the video have been compared, we get the whole candidate key frame sequence.

In the candidate key frame sequence, we also select the first frame as the reference frame, *i.e.*, it is used as the first frame of the endpoint frame sequence. The subsequent candidate key frame, in turn, is compared with the reference-frame in the H-component global cumulative histogram by the formula of the Euclidean distance. When the distance between two frames is greater than or equal to δ, the subsequent frame is used as the second frame of the endpoint frame sequence and it is regarded as a new reference-frame, and so on. We get the whole endpoint frame sequence until the last candidate key frame.

5. **Experimental Results and Analysis.** In order to determine the possible range of the three parameters $(\delta, \delta_1, \delta_2)$, our scheme uses the control variable method to observe experimental results and then gets a fine tradeoff($0.025 \leq \delta \leq 0.045; 35 \leq \delta_1 \leq 45; 10 \leq \delta_2 \leq 20$). Table 1 below shows the result of the comparison between this algorithm and the method using the comparison of the CGCH combined with the comparison of central moment of each original block. Algorithm 1 directly corresponds to the method using the comparison of the CGCH combined with the comparison of central moment of each original block, and algorithm 2 corresponds to this algorithm. Through Table 1, it shows the distinction between algorithm 1 and algorithm 2 intuitivelyAlgorithm 2 is faster than algorithm 1 in dealing with video, and it takes less time to calculate.

Among the seven videos, video3,4 are got by increasing the background frames of video 1,2. So the video duration increases from 50s to 150s, *i.e.*, the number of foreground frames does not change, but the number of background frames increases. The video 1 is compared with the video 3 and the video 2 is compared with the video 4. Comparison of results show that in the same conditionThe number of the foreground frames is the same, the more the background frames, the more advantages of this algorithm.

Video 3-7 are compared with each other. Comparison of results show that in the same conditionThe same video duration, the number of background frames has little effect on algorithm 1. But for algorithm 2, the more the background frame, the shorter the processing time of algorithm. This is very consistent with the characteristics of video surveillance.

The content of the video 5 is more complicated than the content of the video 6, and they almost have the same number of background frames. Through the processing time of algorithm 2, we can see that when the number of background frames is almost the same, the processing time of algorithm 2 will increase for the complicated scene content. But it is still shorter than the processing time of algorithm 1.

Table 1. The comparison of experimental data between algorithm 1 and algorithm 2

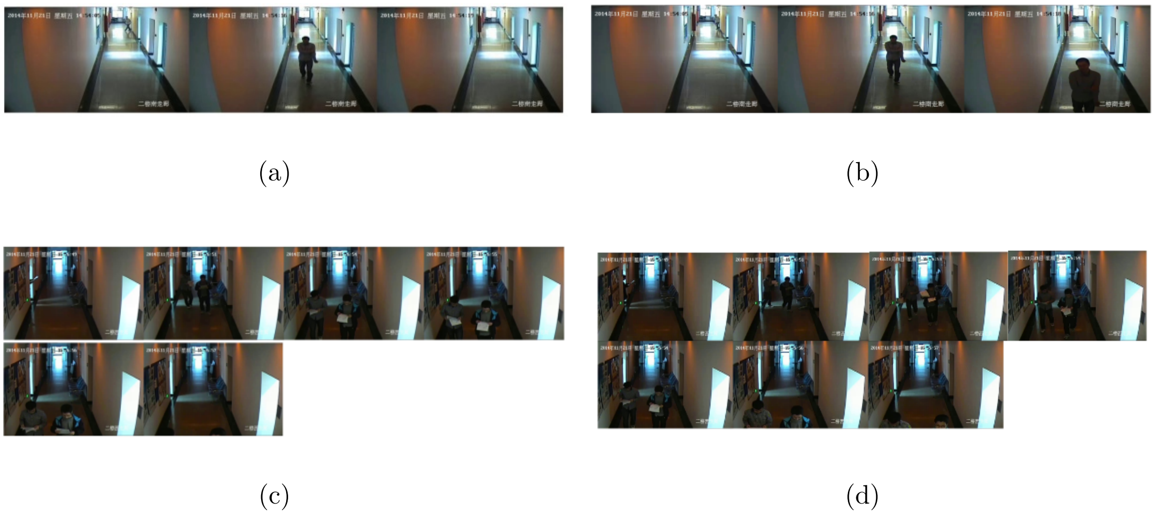
Video Number	1	2	3	4	5	6	7
m(The Number of blocks)	3	3	3	3	3	3	3
The video duration in seconds	50	50	150	150	150	150	150
The algorithm1 in seconds	154	143	452	448	449	447	447
The algorithm2 in seconds	110	98	278	272	374	302	289
The number of key frames by algorithm1	3	6	3	6	32	27	15
The number of key frames by algorithm2	3	7	3	7	26	27	10

The duration of FF(BF) in seconds	12(FF)
	38(BF)
		7(FF)
		43(BF)
			12(FF)
			138(BF)
				7(FF)
				143(BF)
					74(FF)
					76(BF)
						62(FF)
						88(BF)
							16(FF)
							134(BF)

* Note: The algorithm 1: The method using the comparison of the CGCH combined with the comparison of central moment of each original block; The algorithm 2: This algorithm; FF: Foreground frame; BF: Background frame.

Fig.9(a) and 9(b) are the results of key frame extraction for the video 1,3 ((a) corresponds to algorithm 1, (b) corresponds to algorithm 2). Fig.9 (c) and Fig.9(d) are the results of key frame extraction for the video 2,4 ((c) corresponds to algorithm 1, (d) corresponds to algorithm 2). Fig.10(a) and Fig.10(b) are the results of key frame extraction for the video 5 ((a) corresponds to algorithm 1, (b) corresponds to algorithm 2). Fig.11(a) and Fig.11(b) are the results of key frame extraction for the video 6 ((a) corresponds to algorithm 1, (b) corresponds to algorithm 2). Fig.12(a) and Fig.12(b) are the results of key frame extraction for the video 7 ((a) corresponds to algorithm 1, (b) corresponds to algorithm 2). Through comparative observation, we can conclude that the key frames extracted by algorithm 2 are similar to the algorithm 1, but the structure of key frames obtained by algorithm 2 in some video is more compact and evenly distributed.

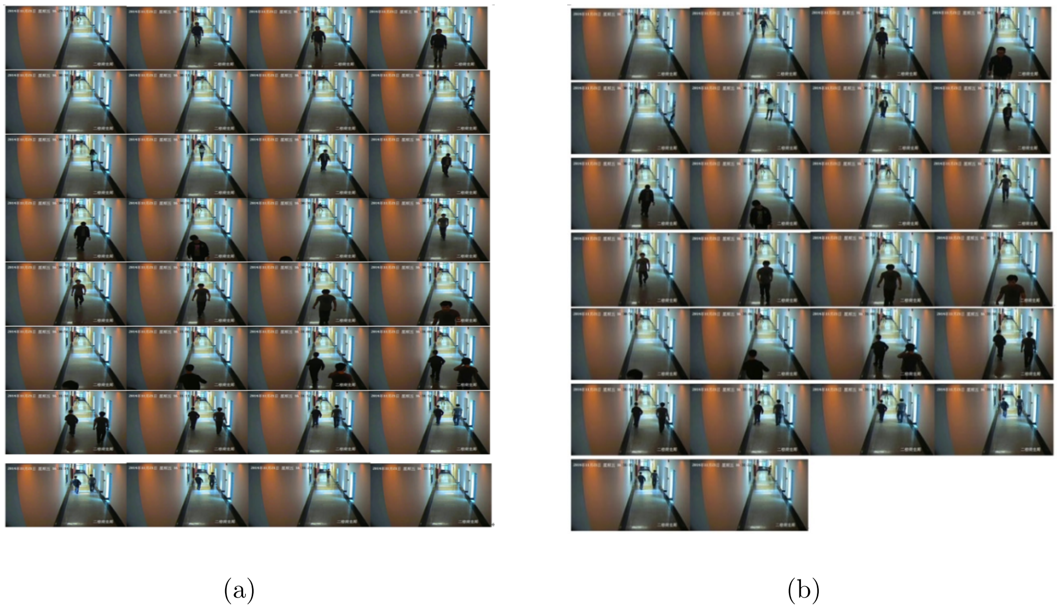
The biggest characteristic of our algorithm is to enhance the speed of the original key frame extraction algorithm specially for surveillance video with a high number of



(a) (b)

(c) (d)

Figure 9. The results of key frame extraction for the videos 1,3 and 2,4. (a) corresponds to Algorithm 1 for videos 1,3; (b) corresponds to Algorithm 2 for videos 1,3; (c) corresponds to Algorithm 1 for videos 2,4; (d) corresponds to Algorithm 2 for videos 2,4

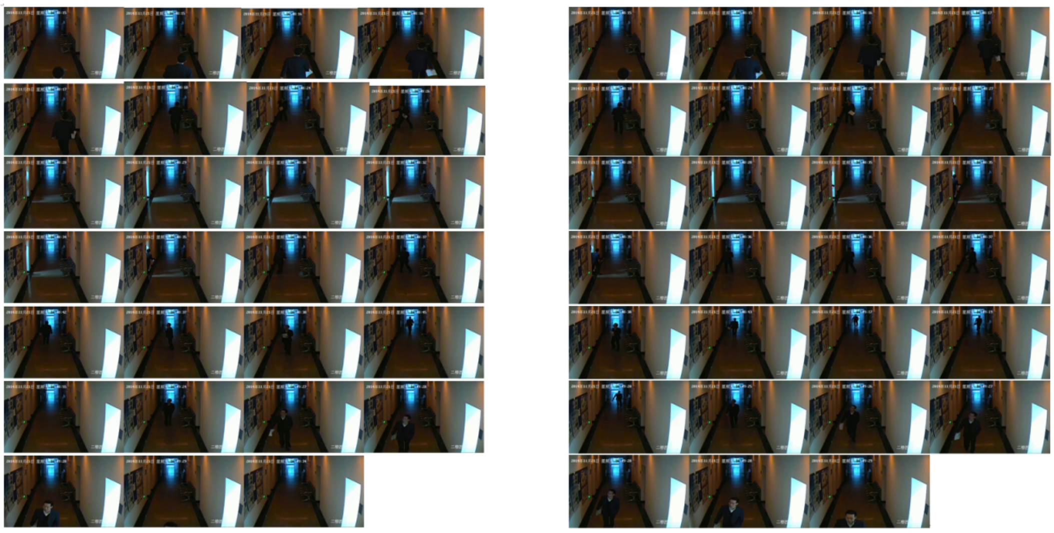


(a) (b)

Figure 10. The results of key frame extraction for the video 5. (a) corresponds to Algorithm 1; (b) corresponds to Algorithm 2.

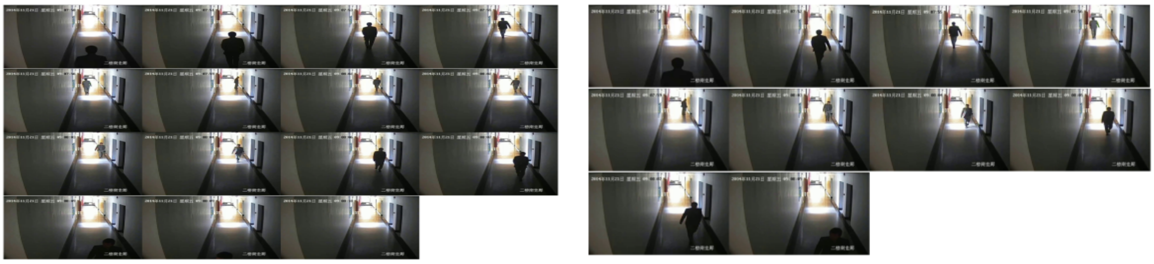
background frames and a simple scene content, and the structure of the key frames is more compact and evenly distributed. But for the surveillance video with the complicated scene content, our algorithm is still faster than the original algorithm, but its processing efficiency will decline.

6. **Conclusions.** In order to increase the processing efficiency of the algorithm using the comparison of the CGCH combined with the comparison of central moment of each original



(a) (b)

Figure 11. The results of key frame extraction for the video 6. (a) corresponds to Algorithm 1; (b) corresponds to Algorithm 2.



(a) (b)

Figure 12. The results of key frame extraction for the video 7. (a) corresponds to Algorithm 1; (b) corresponds to Algorithm 2.

block, our paper proposes a fast key frame extraction method using the GMM foreground extraction combined with the original algorithm. The foreground extracted by GMM algorithm is positioned by the eroding and dilating combined with the 4-neighborhood search algorithm, so that the foreground frames are selected from surveillance video and these frames are rearranged to form a new video sequence. Then the key frames are extracted by the original algorithm. We get the key frames sequence at last. Compared with the original algorithm, this algorithm greatly reduces the number of video frames which are processed by the original algorithm, *i.e.*, greatly reducing the processing time. But in terms of surveillance video, this algorithm still cannot be real-time. However, it provides a basic directionAs long as we find an efficient foreground extraction method combined with a better key frame extraction algorithm, we can greatly reduce the processing time of the key frame extraction algorithm in the surveillance video.

Acknowledgments. This work was partly supported Zhejiang Provincial Natural Science Foundation of China under Grants No. LY15F010003 and No. LY17F030008. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] AMA Stricker, M Orengo. Similarity of Color Images, *Proceedings of SPIE-The International Society for Optical Engineering*, pp. 381-392, 1970.

[2] BF Momin, GB Rupnar. Key frame extraction in surveillance video using correlation, *International Conference on Advanced Communication Control and Computing Technologies. IEEE*, pp. 276-280, 2017.

[3] GGL Priya, S Domnic. Shot based key frame extraction for ecological video indexing and retrieval, *Ecological Informatics*, vol. 23, no. 9, pp. 107-117, 2014.

[4] C Panagiotakis, A Doulamis, G Tziritas. Equivalent key frames selection based on iso-content principles, *IEEE Transactions on Circuits & Systems for Video Technology*, vol. 19, no. 3, pp. 447-451, 2009.

[5] R Vzquez-Martn, A Bandera. Spatio-temporal feature-based key frame detection from video shots using spectral clustering, *Pattern Recognition Letters*, vol. 34, no. 7, pp. 770-779, 2013.

[6] G Ahalya, HM Pandey. Data clustering approaches survey and analysis, *International Conference on Futuristic Trends in Computational Analysis and Knowledge Management*, pp. 532-537, 2015.

[7] C Zhang, Q Xia, G Yang. Reconsideration about clustering analysis, *Industrial Electronics and Applications. IEEE*, pp. 1517-1524, 2015.

[8] SK Kuanar, R Panda, AS Chowdhury. Video key frame extraction through dynamic Delaunay clustering with a structural constraint, *Journal of Visual Communication & Image Representation*, vol. 24, no. 7, pp. 1212-1227, 2013.

[9] DY Kim, H Park. An Efficient Motion-Compensated Frame Interpolation Method Using Temporal Information for High-Resolution Videos, *Journal of Display Technology*, vol. 11, no. 7, pp. 580-588, 2015.

[10] F Kamisli. Recursive Prediction for Joint Spatial and Temporal Prediction in Video Coding, *IEEE Signal Processing Letters*, vol. 21, no. 6, pp. 732-736, 2014.

[11] T Liu, HJ Zhang, F Qi. A novel video key-frame-extraction algorithm based on perceived motion energy model, *IEEE Transactions on Circuits & Systems for Video Technology*, vol. 13, no. 10, pp. 1006-1013, 2003.

[12] Y Ma, Y Chang, H Yuan. Key-frame extraction based on motion acceleration, *Optical Engineering*, vol. 47, no. 9, pp. 957-966, 2008.

[13] C Stauffer, WEL Grimson. Adaptive Background Mixture Models for Real-Time Tracking, *IEEE Computer Society*, vol. 2, pp. 2246, 1999.

[14] T Bouwmans, F El Baf, B Vachon. Background Modeling using Mixture of Gaussians for Foreground Detection: A Survey, *Recent Patents on Computer Science*, vol. 1, no. 3, pp. 219-237, 2008.

[15] DK Yadav. Efficient method for moving object detection in cluttered background using Gaussian Mixture Model, *International Conference on Advances in Computing, Communications and Informatics. IEEE*, pp. 943-948, 2014.