Surveillance Video Retrieval Based on Moving Objects Detection

Huang Xuan [1, 2, 3]
1. Cognitive Science Department, Xiamen University, Xiamen 361005, China
2. Fujian Key Laboratory of the Brain-like Intelligent Systems (Xiamen University), Xiamen 361005, China
3. ZhangZhou institute of technology, ZhangZhou 363000, China

Fernando Croquer
Satya Wacana Christian University, Indonesia
Email: f.dpbgb@yahoo.com

Wenhua Zeng
School of software, Xiamen University, Xiamen 361005, China

Abstract—**In this article, we proposed a surveillance video retrieval method based on detection of moving objects and researched on key problems in surveillance video, such as moving objects detection, video segmentation, selection and description of key frames, and measurement of feature similarities. We detected moving objects via movement detection algorithm based on codebook, and realized automatic segmentation of surveillance video via calculating present Frame Movement Amount based on the foreground and background information in movement detection. After the above process, we proposed a video retrieval method based on the combination of SIFT and color feature.**

Index Terms—**Surveillance Video; Object Detection; Key Frame; SIFT Feature**

I. Introduction

People are more and more concerned about security recently, therefore the demand on intelligent surveillance system increases in some places, such as banks and airports. Meanwhile, the development of mass storage devices ensures the application of surveillance in a variety of occasions. A huge mass of video information brought by the prevalence of surveillance brings us challenges in extracting useful information efficiently. The description of video via traditional text recording is not suitable due to the abundance and variability of surveillance video: for one thing, text recording cannot fully describe the contents of surveillance video; for another, it brings a huge waste of human and material resources. Thus, the problems on the effective description of surveillance video and the efficiency of quick retrieval have become urgent. With the development of computer technology, researchers from the entire world have focus more and more on the content information in surveillance video and a large amount of researches are performed on this topic. Contents-based surveillance video retrieval technique is a newly merged technique with the development of surveillance system and computer vision, which has limited literature resources. The surveillance system has gone through the analog surveillance system before the 1990s, the digital surveillance system in the 1990s and the intelligent surveillance system after the 21st century [1, 14, 15], with the development of network and computer technology. With people more and more concerned on security, intelligent surveillance system will shape the development of surveillance systems and is highly valued because of its wide application perspectives and potential economic values.

Compared with text information, image information features in extensive information contents. On the other hand, the amount of images increases rapidly and the ability to process them is ineffective, therefore large amounts of image data do not come into use. The effective processing of the extensive information contained in images and the efficient retrieval of required images from the huge image database prevail currently. Generally speaking, Content Based Image Retrieval (CBIR) [2] is a technique to solve the above problems. Based on the visual information of images, it can inquiry feature database according to user-defined retrieval requirements and provide users images which meet the similarity criteria. In addition, users can provide feedbacks for the retrieval results, which will result in more accurate outcomes. Simply speaking, Content Based Surveillance Video Retrieval is the process that computers automatically analyze the contents of surveillance video and extract information features to make an index, and the retrieval system automatically returns images or video segments that meet the requirements when a user submits an inquiry about his/her interested contents. This technology consists of detection of moving objects in surveillance video, feature extraction, similarity match and so on. It is involved with image processing, computer vision, pattern recognition and so on, and has become the focus of recent research.

According to the retrieval requirements users submitted, Content Based Video Retrieval can be divided

doi:10.4304/jmm.8.6.762-769

into two technical problems: Content Based Image Retrieval and Content Based Video Segment Retrieval. In the area of image retrieval, the principal method is to match images via extracting low-layer features of images, which is applicable to trademark inquiry, medical image analysis and determination, and registration of remote sensing images. In the area of video segment retrieval, currently application focuses on news video [3] [4] [16] [17] and sports video [5]. Generally speaking, currently developed systems are mainly designed for news video and sports video, and researches on surveillance video are scarce. Researchers at home and abroad have made some achievements on surveillance video retrieval. For example a video surveillance project VSAM, which is conducted by USA Defense Advanced Research Projects Agency, has realized intelligent surveillance on future battlefield and moving vehicles as well as pedestrians in city traffic. W4 System developed by University of Maryland can basically realize real-time tracking of pedestrians and predict their movements when the moving objects of surveillance video are pedestrians. Compared to studies abroad, studies on Content Based Video Retrieval at home started late and reside at initial edge. A large number of institutes have stepped into this research area and made some achievements. For example, TV-FI (Tsinghua Video Find It) organized and developed by Tsinghua University has realized entry management of video database and provides content based browse as well as retrieval. Web scope-CBVR [6] system developed by Zhejiang University provides inquiry method based on keywords and video examples: when a user inquiries via this system, he/she can provide either keywords that contain the theme of the video or a segment of video example, then the system will automatically return the video that most satisfies the retrieval requirements. In addition, institutes such as Microsoft Research Asia, Shanghai Jiaotong University, Fudan University and National University of Defense Technology are performing related researches and have developed some experimental systems. The researcher team led by Professor Ziqing Li has achieved great progress on intelligent surveillance video. Their system can detect and track of pedestrians and vehicles in surveillance video as well as recognize abnormal behaviors and send alert.

The traditional surveillance video retrieval is based on key words which are labeled by people such as traditional image retrieval. This way produces many drawbacks such as: time consuming, uncertainty and semantic gap and so on. Surveillance video retrieval system based on this way also makes use of other attributions such as system times. In order to retrieve a large length video, you must know the exact time. But it is very pity as to the surveillance video the most people can not provide the time. The intelligent surveillance comes with the developments of CBIR. CBIR is greatly developed in the recently years, which has been used in Baidu and Google. The surveillance video retrieval based on CBIR also makes great processions. The system extracts image globe features such as color, texture or sharp, then constructs index to search the video database. Although the methods are simple but the globe features are not well suitable to represent the image contents. Local features such as SIFT are robust to scale, rotation, light and occlusion. So local features extracted from images are used to represent images, which has been proved successful in image retrieval field such as medical image retrieval, landmark image retrieval and so on. However, in accurate application, the users are not only interest in image retrieval but also in video segment retrieval, which provides more information and makes surveillance video more useful. As we all know, the surveillance video has too much no useful information, which is determined by its characters. In order to improve the retrieval efficient of surveillance video, it is not wise to take the same methods for news video and sports video retrieval. In this paper, we present a new algorithm to solve the problems of surveillance video. In order to improve surveillance video retrieval efficient, moving objects are detected based on codebook method, which is shown more accurate than other ways from experiments. We classify each frame according to its movement amount and surveillance video is segmented according to the class borders of its frames. In traditional retrieval systems, the globe color feature such as color histogram is used for image retrieval and recently the local SIFT feature is extracted for image represent. Image retrieval based on the two features all make great successful. In this paper, the SIFT feature and color feature are combined, which achieves more accurate retrieval results.

This work researches on and gives reasonable solutions for key problems in surveillance video, such as moving objects detection, video segmentation, selection and description of key frames, and measurement of feature similarities. First of all, we used codebook based movement detection algorithm to detect moving objects. Then, we calculated the movement amount of current frame mainly based on the foreground and background information in movement detection. Thus, automatic segmentation of surveillance video is achieved. In addition, by including SIFT (Scale Invariant Feature Transform) feature in surveillance video retrieval, we proposed a video segment retrieval algorithm based on the combination of SIFT and color feature. At last, we proved the effectiveness of the above algorithms via a large amount of experiments.

II. SURVEILLANCE VIDEO RETRIEVAL BASED ON DETECTION OF MOVING OBJECTS

A. Moving Object Detection Based on Codebook

In real life traffic, surveillance system is applied in more and more occasions. Thanks to the development of storage techniques such as mass storage devices and DVR, these surveillance video can be preserved. On the other hand, with the enhancement of concern on safety, more and more people have paid attention to traffic problems such as vehicles violating traffic rule and hit-and-run. After the obtainment of video segments or images of target vehicles, it is hoped that all video segments containing these vehicles can be retrieved efficiently. Moving objects detection is the key problem of traffic surveillance video retrieval. It is the basis of further researches and has a direct influence on the results of experiments. Considering the frequent background changes in surveillance video, such as change of light and vibrancy of leaves, we mainly detected moving objects based on codebook [7]. This method can achieve satisfactory experimental results when detecting moving objects in complicated background. Its basic procedure is as following:

1) Generation of Background Codebook Model

Background codebook is generated from the continuous sampled value of each pixel. The sampled value of a pixel is presented as X = {$\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_N$}, in which xi (1≤i≤N)is a vector in the color space. In this article, we used the RGE color space. The codebook of a pixel is presented as C = {$\boldsymbol{c}_1$, $\boldsymbol{c}_2$, ..., $\boldsymbol{c}_N$}. A code letter $\boldsymbol{y}_k$ in the codebook is a binary <$\boldsymbol{U}_k$, $\boldsymbol{V}_k$>, in which $\boldsymbol{U}_k = (R, G, B)$ and $\boldsymbol{V}_k = (b_k^{\min}, b_k^{\max}, f_k, \lambda_k, p_k, q_k)$, $b_k^{\min}$ and $b_k^{\max}$ are respectively the minimum and maximum value of the pixel intensity of the code letter, f_k is the frequency that the code letter appears, λ_k is the maximum interval between two successive appearance of the code letter, and p_k and q_k are respectively the first time and the last time that the code letter appears. For the value of each pixel in a given training video $\boldsymbol{x}_t = (R_t, G_t, B_t)$, if the codebook is empty, then M = 0. Generate a new code letter as following:

$$M = M + 1, b = \sqrt{R_t^2 + G_t^2 + B_t^2} \tag{1}$$

$$U_M = (R, G, B), V_M = (b, b, 1, t-1, t, t) \tag{2}$$

If the codebook is not empty and the following formula (3) and (4) are satisfied, find code letter corresponding to $\boldsymbol{x}_t$ so that $\boldsymbol{x}_k$ ($\boldsymbol{U}_k = (R_k, G_k, B_k)$, $\boldsymbol{V}_k = (b_k^{\min}, b_k^{\max}, f_k, \lambda_k, p_k, q_k)$. Formula (3) and (4) are:

$$colordist(x_t, x_k) \le \varepsilon \tag{3}$$

$$brightness(b, < b_k^{\min}, b_k^{\max} >) = true \tag{4}$$

in which the distance between colors *colordist* and the border of brightness are defined as:

$$colordist(x, y) = 1 - \sqrt{\frac{(R_x R_y + G_x G_y + B_x B_y)^2}{(R_x^2 + G_x^2 + B_x^2)(R_y^2 + G_y^2 + B_y^2)}} \tag{5}$$

$$brightness(b, < b_k^{\min}, b_k^{\max} >) = \begin{cases} true, & b_i^{\min} \le b \le b_i^{\max} \\ false, & otherwise \end{cases} \tag{6}$$

(R_x, G_x, B_x) and (R_y, G_y, B_y) are the color vectors of any two pixels. After finding the corresponding code letter, update the color letter according to:

$$U_k = (\frac{f_k R_k + R_t}{f_k + 1}, \frac{f_k G_k + G_t}{f_k + 1}, \frac{f_k B_k + B_t}{f_k + 1}) \tag{7}$$

$$V_M = (\min\{b, b_k^{\min}\}, \max\{b, b_k^{\max}\}, f_k + 1, \max\{\lambda, t-q\}, p_k, t) \tag{8}$$

After the training of given video, we calculate λ the maximum interval between two successive appearance of the code letter and delete the redundant code letters, so that the initial codebook that can simulate the actual background.

2) Detection of Foreground Moving Objects

For newly input pixel $\boldsymbol{x} = (R, G, B)$, define a Boole variable *mflag* and set its initial value to 0. The corresponding codebook is M. First of all, its brightness is calculated according to formula (1). Then $\boldsymbol{x}_k$, the corresponding code letter to $\boldsymbol{x}$ is searched in M. If it is found, then *mflag* = 1. At last, the pixel region of foreground moving objects is determined according to:

$$B = \{ \begin{matrix} foreground, & mflag = 1 \\ background, & otherwise \end{matrix} \tag{9}$$

The effects of this method and other methods are shown in Figure 1, Figure 2. In these figures, Image (a) is the current frame in the surveillance video during movement detection. Image (b) is the impressing drawing of movement detection when applying Three-frame Difference Method [8], Image (c) Mixed Gaussian Back Ground Model [9] and Image (d) method in this article.

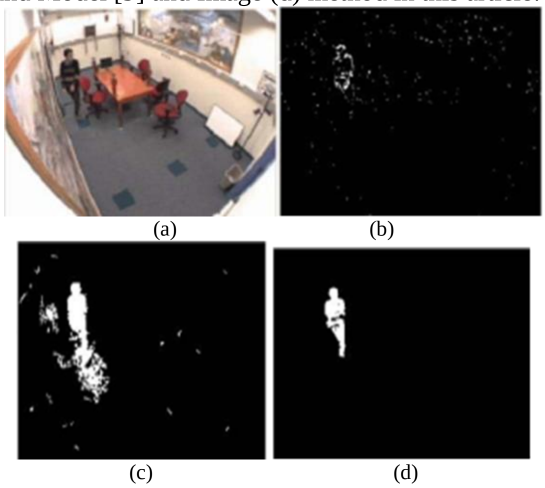

(a) (b)
(c) (d)

Figure 1. Comparison of video motion detection of indoor surveillance

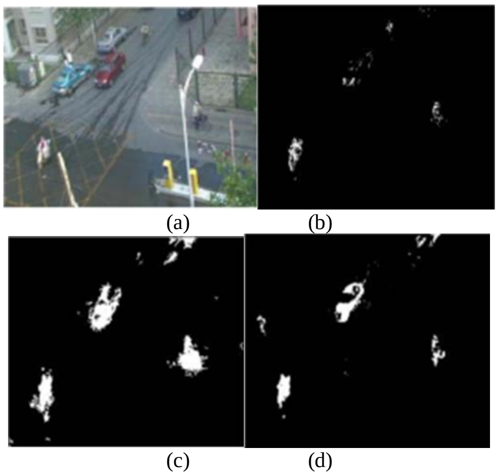

(a) (b)
(c) (d)

Figure 2. Comparison of video motion detection of complex background surveillance

B. Video Segmentation

Surveillance video segment consists of a series of continuous frames with interrelated lexeme and describing the same event. From the users' view, the inquiry to surveillance video database includes a video segment containing one or more user-interested events, so that the results are meaningful, for example, the appearance of a vehicle in the surveillance system, the search for a target vehicle in the surveillance video database and the judgment whether a vehicle appeared in the surveillance video of a certain day. Based on such consideration, surveillance video must firstly be segmented to satisfy the need of users to inquiry via video segments. Because of the fixed position of surveillance devices and its limited coverage, the contents of surveillance are mainly the movement of pedestrians and vehicles, and the change of surveillance scene is limited, typically the gradual or sudden change of light. In a surveillance video, the amount and rate of changes of surveillance frames reflect the amount and movement mode of moving objects in the surveillance scene. According to this feature, we classify surveillance video via Frame Movement Amount [11], and segmented it eventually.

Video segmentation mainly includes background establishment and update, the calculation of Frame Movement Amount and frame classification.

1) Background Establishment

The selection of background frame is mainly based on the immobile objects in video database and certain frame can be selected as the background frame. However, considering the possibility and feasibility that there is no moving object in a complex of background, we may be not able to obtain a background image without moving objects. In this case, we can obtain the background frame at the moment that there are only a few moving objects. Specifically, we can select a long period for a single pixel in continuous frames and make the median of brightness the brightness of the pixel in the initial background frame. The median of brightness is calculated as:

$$B(x,y)=\frac{1}{N}\sum_{i=0}^{N-1}F_i(x,y) \qquad (10)$$

In which $B(x, y)$ is the initial background and $F_i(x, y)$ is a frame in the surveillance video.

2) Background update and Frame Movement Amount

Frame Movement Amount is defined as the ratio of the number of pixels in the moving area to the total number of pixels in the frame. It reflects the occupancy of moving objects in the surveillance video. Obviously, its value is between 0 and 1. Let the size of a frame is $h \times w$, $C_t(x, y)$ is the brightness at pixel point (x, y) in the current frame, and D_t is binary image of corresponding frame difference, then:

$$D_t(x,y)\begin{cases}0, & |C_t(x,y)-B_t(x,y)|<T\\ 1, & |C_t(x,y)-B_t(x,y)|\geq T\end{cases} \qquad (11)$$

Considering the influences of noise from the environment, the binary image of frame difference may has null value in moving objects. Thus, we can further process the binary image via Morphological Close Operation and Connected Domain Combination, until the movement regions of actual moving objects are determined. Frame Movement Amount M_t can be obtain from the binary image according to:

$$M_t=\frac{1}{h\times w}\sum_{x=1}^{h}\sum_{y=1}^{w}D_t(x,y) \qquad (12)$$

3) Classification

Divide video frames in to 6 different classes according to Frame Movement Amount M_t. Label the current frame in the surveillance video, for example the class of mth frame is S_k, whose value is an integer between 0 and 5.

TABLE I. CLASSIFICATION OF FRAMES

Rage of Frame Movement Amount M_t	Label of Class S_k
$0 \leq M_t < 0.1$	$S_0 = 0$
$0.1 \leq M_t < 0.3$	$S_1 = 1$
$0.3 \leq M_t < 0.5$	$S_2 = 2$
$0.5 \leq M_t < 0.7$	$S_3 = 3$
$0.7 \leq M_t < 0.9$	$S_4 = 4$
$0.9 \leq M_t < 1$	$S_5 = 5$

4) Classification of Video

Via the comparison between S_k and S_{k-1}, i. e. the classes of the current and the previous frame, we can locate the border of different classes. By labeling and classifying frames, we can build up a hierarchical structure of the whole video, in which different classes are not independent, but rather we can locate the borders between them according to their time sequences, i. e. the starting or end frames of classes. Then we can segment the video via starting and end frames as following:

If $S_{k-1} = S_k$, then the two frames are of the same class and without border, so we can move on to the next frame;

If $S_{k-1} < S_k$, then k is the starting frame of the class from S_{k-1} +1 to S_k;

If $S_{k-1} > S_k$, then k – 1 is the end frame of the class from S_k +1 to S_{k-1}.

C. Video Segmentation Retrieval based on SIFT

After segmentation, the contents of a video segment can be presented by some key frames, so that the redundancy of the video database is greatly reduced.

1) Key Frame Selection

Key frame is the image frame that is representative in a series of images and can reflect the general contents of the video segment. Via the video segmentation method in Section 2. 2, surveillance video is divided into basic units within which the contents are similar. To proceed, we can select the key frames from these units and make them represent the contents of the basic units. Because a surveillance video is shot under a fixed occasion, the scenes recorded are almost the same and there is a large amount of redundant information in the frames of the video segments. Thus, the selection of key frames can effectively keep the main contents of the video meanwhile greatly reduce the amount of information contained in this video. As a result, the video segment retrieval rate is improved dramatically.

Generally, the key frame selection is based on two purposes: first of all, it is able to present the theme and part of the contents of the video; in addition, it can change the research on dynamic video into that on static images, which simplifies the research as well as extracts the features of the key frame, such as color, pattern and shape, and makes them the date sources of video abstract database index. According to the feature of surveillance video, the difference between two segments centers on the differences of moving objects and the brightness of light. Surveillance video is segmented according to Movement Amount. Here we apply a more simple method to select key frame, i. e. selecting the starting frame of a video segment as its key frame.

2) SIFT Feature Extraction

The selection of features directly influences the eventual results. The characteristics of different features are: color is a global feature which can well describe the color properties of moving objects in the surveillance video; patter is also a global feature, but considering the environmental influences in the video such as light and illusionary patterns will be caused by the influence of light, the result will be disturbed; The shapes and spatial relations of moving objects are greatly influenced by the division of images, meanwhile spatial relations are sensitive to the transportation and rotation of moving objects, so they cannot be used in the retrieval of surveillance video. Based on the above consideration, here we mainly apply the combination of SIFT and color feature to describe the contents of images.

SIFT [12] was proposed by David Lowe in 2004 when he was summarizing testing methods of non-variable technique. It is an operator describing the local properties of images such as measurement space, image scaling, rotation and even affine invariant. SIFT is adaptive to the change of light and the rotation and twisting of images. The essence of applying SIFT is to extract SIFT Key Points from images. SIFT feature extraction includes four steps:

a) Detection of Extremes in Measurement Space

The introduction of Measurement Space Theory provides better simulation of the multi-scale feature of image data. Koendetink [12] has proved that Gaussian Convolution Kernel is the only transformation kernel to perform scale transformation. Further study of Lindeberg [13] etc. shows that it is also the only linear kernel. To improve the specificity and robustness of the feature, Lowe detected the extremes in both two-dimensional space and DOG (Difference of Gaussian) Space and set them as characteristic points.

b) Determination of Key Points: in Step

We have got all the potential key points. These points must be sharply distinguished from its surrounding regions, i. e. key points should not be low contrast, and meanwhile it should not be edge point. During the process of key point determination, we fit three-dimensional quadratic functions to accurately calculate the locations and scales of key points and meanwhile delete the low contrast or edge points.

c) Determination of the sizes and directions of key points

Determination of the principal direction of a key point via gradient histogram ensures the rotational invariant of the operator. Sampling is performed in the neighboring region whose center is a key point, and then histogram is used to summarize the gradient directions of this region. The scale of gradient histogram is 0 to 360 and each column represents 10 degrees, so that the histogram contains 36 columns. The height of a column in the histogram represents the magnitude of gradient of neighboring region in that direction. After comparison of the magnitudes of all columns, the gradient corresponding to the peak value is the principal direction of the neighboring region of this key point, i. e. the direction of the key point.

d) Generation of SIFT Eigenvector Descriptor:

After obtaining the direction of a key point, we will generate SIFT Eigenvector Descriptor. SIFT Eigenvector is also generated from the statistical information of the gradient in the neighboring region of the key point. To make sure the rotational invariant of SIFT Eigenvector, firstly we rotate the axes to the direction of the key point. According to Lowe's method, we select a 16 × 16 window centering on the key point. Then we divide the window to 4 × 4 squares and make the gradient histogram of each square with 8 directions. Therefore, a characteristic point consists of 16 seed points and each seed point has the vector information of 8 directions. These 16 seed points have altogether the vector information of 128 directions, which are combined into a 128-dimensional vector and the vector is the descriptor of the key point.

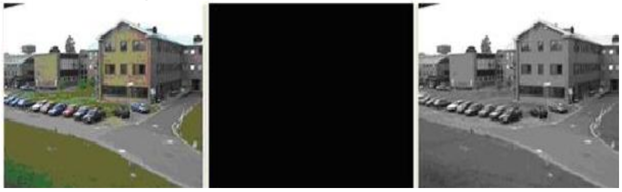
3) Similarity Measure

The retrieval of video segments can be mathematically described as the following: the video sequence with q frames provided by a user is represented as $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_q\}$, and its descriptor is F($\mathbf{X}$). The video sequence with p frames in the video database is represented as $\mathbf{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_p\}$. Generally, users provide a part of video segments to retrieve all similar segments from the video database, because the segment to be retrieved is usually shorter than the video database. Perform video classification on Y and measure the similarity between each segment to the inquiry video. During the process of video segment retrieval, the measurement of the similarity between different video segments is the critical step which directly influences the accuracy of the retrieval result. Such measurement is the similarity measure. The measurement of similarity is typically the measurement of the distance between the features of key frames and then determining the similarity according the distance. Frequently used distances are Euclidean distance, absolute value distance, functional distance and similarity coefficient of extrernes. Here we mainly measure the similarity between relevant video segments according to Euclidean distance.

III. Experiments

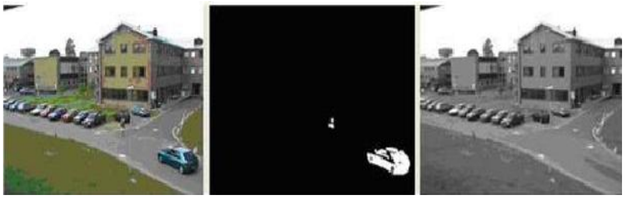
A. Background Detection and Key Frame Selection

In this experiment, we used a segment of the segmented surveillance video as the input segment, which is shown in Figure 3. Perform moving object detection on the input segment, and select the current frame as the key frame when it is the first time that the complete moving object appears in the surveillance image. The key frame is shown in Figure 4. Meanwhile, select the key frame as the input image for SIFT feature extraction.



(a) Original video (b) Foreground image (c) Background image

Figure 3. Initial video sequences



(a) Original video (b) Foreground image (c) background image

Figure 4. Motion detection image

When the moving object appears on the surveillance video, the system called movement detection module to perform movement detection. The impression drawing of foreground and background detection of the surveillance video is shown in Figure 4.

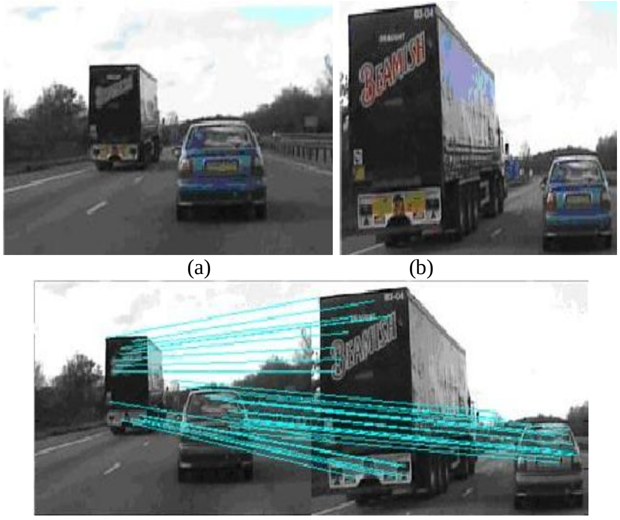
The key frame selected after video segmentation is shown in Figure 5. Because surveillance video retrieval was mainly focused on moving objects, we mainly extracted SIFT features of the moving object which is shown in Figure 6.

Figure 5. Key frame

Figure 6. Moving target

B. SIFT Extraction and Match

Key points should be detected before the extraction of SIFT feature. Compare the value of each pixel in DOG image with its 26 neighboring pixels. If a pixel was an extreme point, it was selected as a potential key point. Then we determined the size and direction of a key point via gradient histogram. In the same process, we firstly classified the user provide video database and measured the similarity between every segment in this database and the segment to be retrieved. Take one of the segments as an example. The moving object detected from this segment is shown in Figure 6. Extract the moving object's SIFT features of the video segment to be retrieved and every segment in the database in to 128-dimensional vectors. Then calculate the similarity of SIFTs between two images according to their Euclidean distance. In this experiment, for each key point in the key frame of inquiry segment, calculate its distances from two nearest key points in the video database d1, d2. If d1/d2 < Th, then the two key points are a pair of matching points. Usually the value of Th is 0. 6. The image matching result in this experiment is shown as following in figure 7: Image (a) is an image containing the moving object extracted from the inquiry segment; Image (b) is an image containing the moving object extracted from the video database; Image (c) is the result of matching between the inquiry segment and the video database.



(a) (b)

(c)

Figure 7. Image matching

Whether two images are similar can be determined by the matching result of their key points. The more matching key points in two images, the more similar they are. Usually we define a distance to measure the similarity of two images, which is:

$$dist = 1 - \frac{f_{qp}}{\min(f_q, f_p)} \tag{13}$$

in which fq is the number of key points in the inquiry image and fp in the compared image. fpq is the number of matching key points. The above equation show that the smaller the distance is, the similar the two images are. Calculate the distances between the inquiry segment and every segment from the video database and sort the distances. Select six segments with the smallest distances as the retrieval result.

C. Video Segment Retrieval

The video database used in this experiment was combined from PETS surveillance video database, surveillance video provided by companies and video recorded by ourselves. For PETS, mainly PETS2000 and PETS2001 video database were selected, and related contents of FETS video database can be downloaded from the ftp server of University of Reading. Surveillance videos provided by companies were mainly recorded by actual traffic surveillance system and district surveillance system. Video recorded by ourselves were campus surveillance video. The scenes in the video database included indoor porch, residential districts, campus and highway. We extracted 160 video segments as inquiry segments, and meanwhile included 10 segments from other video as inquiry segments. The duration of inquiry segments were about 30s. The retrieval result is shown in Table 2.

TABLE II. THE RESULT OF VIDEO SEGMENT RETRIEVAL

Inquiry video segment	Number of retrieval results	Number of successful results	System detection omission	Recall ratio
Actual traffic surveillance video	42	36	6	85. 8%
PETS surveillance video	37	37	0	100%
Traffic surveillance video recorded by ourselves	36	33	3	91. 7%
Outdoor district surveillance video	26	25	1	96. 20%
Outdoor corridor surveillance video	6	6	0	100%
Hospital surveillance video	23	23	0	100%

The criteria for a successful result is: if the video library contains the same segment as the inquiry segment, then the successful result will return video segment with error less than 5s; if the video library does not contain such a segment, then the successful result will return no information. When the video library contains the same segment as the inquiry segment but the system cannot find it, such situation is called system detection omission. Table 2 shows that this algorithm can maintain a recall ratio of about 85% for actual traffic surveillance video, and for situations with simpler surveillance image, such as district surveillance, hospital surveillance and traffic surveillance video recorded by ourselves, this algorithm has better performance.

IV. CONCLUSION

In this article, we forwarded a method that is based on content based surveillance video segment retrieval and combines SIFT and color features to retrieve video segment from surveillance video. We detected moving objects based on codebook. Then we classified each frame according to its Movement Amount and segmented surveillance video according to the class borders of its frames. Finally, we proposed a feature extraction algorithm to extract the features of key frames in the retrieved segment, and meanwhile measure its similarity to the inquiry surveillance video segment. Sort and return the retrieval result according to similarity measurement. The effectiveness of the video segment retrieval algorithm in this article was tested via a large.

ACKNOWLEDGMENT

This work was partially supported by National Nature Science Foundation of China (No. 61202143), the Nature Science Foundation of Fujian Province (No. 2011J01367), Xiamen University 985 Project and research funds of Zhangzhou institute of technology (ZZY1307).

REFERENCES

[1] Boult Terrance, Johnson RC, Pietre Tracy, et al, “A Decande of Networked Intelligent Video Surveillance”, *In Proceedings of Workshop on Distributed Smart Cameras, Boulder*, USA, 2006.

[2] H. J. Zhang, J. H. Wu, D. Zhongctal, “An integrated system for content-based video retrieval and browsing”, *Pattern Recogition*, vol. 30, pp. 64-655, 1997.

[3] Yuan Junsong, Tian Qi, “Ranganath Surendra, Fast and robust search method for short video clips from large video collection”, *In: Proceedings of the 17th International Conference on Parrern Recognition*, pp. 866 – 869, 2004.

[4] Lienhart R, et al, “On the detection and recognition of television commercials”, *In: Proceedings of IEEE Conference on Multimedia Computing and Systems,* pp. 509-516, 1997.

[5] AnilJain, Aditya Vailaya, Xiong Wei, “Query by video clip”, *Multimedia system*, vol. 7, no. 5, pp. 369-384, 1999.

[6] Zhuang Y, Liu X, Pan Y, “Web scope-CBVR: A customized content-based Search Engine for video on WWW”, *In: Proceeding of IS&T and SPIE Image and Video Communications and Processing*, 2000.

[7] Kim K, Harwood D, et al, “Background modeling and subtraction by Codebook construction”, *Proceedings of IEEE International Conference on Image Processing, Singapore*, pp. 3061-3064, 2004.

[8] N. V. Patel, LK. Sethi, “Video shot detection and characterization for video data bases”, *Patten Reeognition*, vol. 30. , pp. 583- 592, 1997.

[9] Stauffer C, Grimson WEL, “Adaptive background mixture models for real-time tracking”, *In Proceedings of IEEE Conference on Computer Vision and pattern Recognition,* pp. 246-252, 1999.

[10] WaetlarHD, KanadeT, etal, “Intelligent access to digital video: information Project”, *IEEE Computer*, vol. 29, no. 5, pp. 46-52, 1996.

[11] S. H. Kim, R-H. Park, “Robust video indexing for video sequences with complex brightness variations”, *In Proceedings of lnt. Conf. Signal and Image Processing,* pp. 410-414, 2002.

[12] David Lowe, “Distinctive image features from scale-invariant key points”, I*nternational Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[13] Lindeberg T, “Scale-space theory: A basic tool for analyzing structures at different scales”, *Journal of Applied Statistics,* vol. 21, no. 2, pp. 224-270, 1994.

[14] Chen, Yi-Ling ; Chen, Tse-Shih et al. , "Intelligent Urban Video Surveillance System for Automatic Vehicle Detection and Tracking in Clouds ", *2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA),* pp. 814 - 821, 2013

[15] Hae-Min Moon; Sung Bum Pan, "A New Human Identification Method for Intelligent Video Surveillance System", *2010 Proceedings of 19th International*

Conference on Computer Communications and Networks (ICCCN), pp. 1-6, 2010.

[16] Wen-Pinn Fang, "An Intelligent Hand Gesture Extraction and Recognition System for Home Care Application", *2012 Sixth International Conference on Genetic and Evolutionary Computing (ICGEC),* pp. 457 - 459, 2012.

[17] Banic, N, "Detection of commercials in video content based on logo presence without its prior knowledge ", *2012 Proceedings of the 35th International Convention MIPRO,* pp. 1713-1718, 2012.