J Intell Robot Syst (2009) 55:403–421
DOI 10.1007/s10846-008-9297-3

UNMANNED SYSTEMS PAPER

Surveillance Robot Utilizing Video and Audio Information

Xinyu Wu · Haitao Gong · Pei Chen · Zhi Zhong · Yangsheng Xu

Received: 9 April 2008 / Accepted: 12 November 2008 / Published online: 9 January 2009

Abstract For the aging population, surveillance in household environments has become more and more important. In this paper, we present a household robot that can detect abnormal events by utilizing video and audio information. In our approach, moving targets can be detected by the robot using a passive acoustic location device. The robot then tracks the targets by employing a particle filter algorithm. To adapt to different lighting conditions, the target model is updated regularly based on an update mechanism. To ensure robust tracking, the robot detects abnormal human behavior by tracking the upper body of a person. For audio surveillance, Mel frequency cepstral coefficients (MFCC) is used to extract features from audio information. Those features are input to a support vector machine classifier for analysis. Experimental results show that the robot can detect abnormal behavior such as "falling down" and "running". Also, a 88.17% accuracy rate is achieved in the detection of abnormal audio information like "crying", "groan", and "gun shooting". To lower the false alarms by abnormal sound detection system, the passive acoustic location device directs the robot to the scene where abnormal events occur and the robot can employ its camera to further confirm the occurrence of the events. At last, the robot will send the image captured by the robot to the mobile phone of master.

Keywords Surveillance robot · Video surveillance · Audio surveillance

X. Wu (✉) · Z. Zhong · Y. Xu
Department of Mechanical and Automation Engineering,
The Chinese University of Hong Kong, Hong Kong, China
e-mail: xywu@mae.cuhk.edu.hk

X. Wu · H. Gong · P. Chen · Y. Xu
Shenzhen Institute of Advanced Integration Technology,
Chinese Academy of Sciences/The Chinese
University of Hong Kong, Shenzhen, China

1 Introduction

There are many global issues that could be eased by the correct usage of video surveillance, most of which occur in public places such as elevators, banks, airports, and public squares. However, surveillance could also be used in the home. With increasing numbers of aged people living alone, household surveillance system could be used to help the elderly live more safely. The fundamental problem with surveillance systems is the intelligent interpretation of human events in real-time. In this paper, we present a household surveillance robot combining video and audio surveillance that can detect abnormal events.

The testing prototype of the surveillance robot is composed of a pan/tilt camera platform with 2 cameras and a robot platform (Fig. 1). One camera is employed to track target and detect abnormal behavior. The other is planned to detect face and realize face recognition, which is not discussed in this paper. Our robot can first detect a moving target by sound localization and then track it across a large field of vision using a pan/tilt camera platform. It can detect abnormal behavior in a cluttered environment, such as a person suddenly running or falling down to the floor. By teaching the robot the difference between normal and abnormal sound information, the computational action models built inside the trained support vector machines can automatically identify whether newly received audio information is normal. If abnormal audio information is detected, then the robot can employ its camera to further check the events directed by the passive acoustic location device.

A number of video surveillance systems for detecting and tracking multiple people have been developed, such as W^4 in [1], TI's system in [2], and the system in [3]. Occlusion is a significant obstacle for such systems and good tracking often depends on correct segmentation. Furthermore, none of these systems is designed to detect abnormal behavior as their main function. Radhakrishnan et al. [4] presented a systematic framework for the detection of abnormal sounds that may occur in elevators.

Luo [5] built a security robot that can detect a dangerous situation and provide a timely alert, focusing on fire detection, power detection, and intruder detection. Nikos [6] presented a decision-theoretic strategy for surveillance as a first step toward automating the planning of the movements of an autonomous surveillance robot.

The overview of the system is shown in Fig. 2. In the initialization stage, two methods are employed to detect moving objects. One is to pan the camera step by step and employ the frame differencing method to detect moving targets during the static stage. The other method uses passive acoustic location device to direct the camera at the moving object, keeping the camera static and employing the frame differencing method to detect foreground pixels. The foreground pixels are then clustered into labels and the center of each label is calculated as the target feature, which is used to measure the similarity in the particle filter tracking. In the tracking process, the robot camera tracks the moving target using a particle filter tracking algorithm, and updates the tracking target model at appropriate time. To detect abnormal behavior, upper body (which is more rigid) tracking is implemented that uses the vertical position and speed of the target. At the same time, with the help of a learning algorithm, the robot can detect abnormal audio information, such as crying or groaning, even in other rooms.

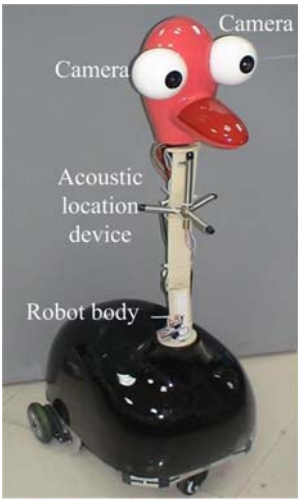

Fig. 1 The testing prototype of surveillance robot

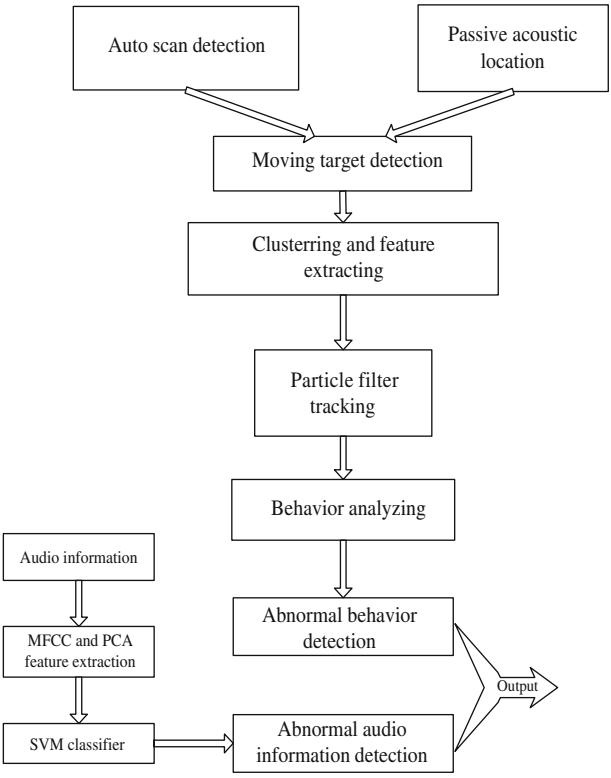

Fig. 2 Block diagram of the system

The rest of this paper is organized as follows. Section 2 introduces the passive acoustic location device, frame differencing method and feature selection and segmentation. Section 3 presents how to employ the particle filter algorithm to track a moving target and detect abnormal behaviors. In Section 4, we present how to employ support vector machine to detect abnormal audio information. Section 5 shows the experimental results utilizing video and audio information before we conclude in Section 6.

2 System Initialization

In many surveillance systems, the background subtraction method is used to find the background model of an image so that moving objects in the foreground can be detected by simply subtracting the background from the frames. However, our surveillance robot cannot use this method because the camera is mobile and we must therefore use a slightly different approach. When a person speaks or makes noise, we can locate the position of the person with a passive acoustic location device and rotate the camera to the correct direction. The frame differencing method is then employed to detect movement. If the passive acoustic location device does not detect any sound, then the surveillance robot turns the camera 30 degrees and employs the differencing method to detect moving targets. If the robot does not detect any moving targets, then the process is repeated until the robot finds a moving target or the passive acoustic device gives it a location signal.

2.1 Passive Acoustic Location

An object producing an acoustic wave is located and identified by the passive acoustic location device. Figure 3 shows the device which comprises four microphones installed in an array. The device uses the time-delay estimation method, which is based on the time differences in sound reaching the various microphones in the sensor array. The acoustic source position is then calculated from the time-delays and the geometric position of the microphones. To obtain this spatial information, three independent time-delays are needed, and therefore the four microphones are set at different positions on the plane. Once the direction result has been obtained,

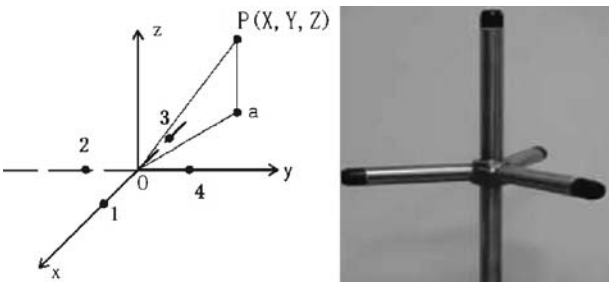

Fig. 3 Passive acoustic location device

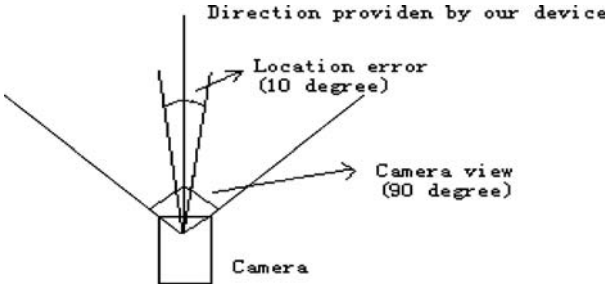

Fig. 4 The method to solve location error

the pan/tilt platform will move so that the moving object is included in the camera view field.

The precision of the passive acoustic location device depends on the distances between the microphones and the precision of the time-delays. We test the device and passive acoustic location error is about 10 degree in the x–y plane. See Fig. 4, the camera angle is about 90 degree and much larger than the passive acoustic location error. When the passive acoustic location device provides a direction, the robot turns the camera and keeps the direction in the center of the camera view. Thus the location error could be ignored.

2.2 Target Detect Using Frame Differencing Method

We employ the frame differencing method to detect a target, as this only requires the camera to be kept static for a while. Frame differencing is the simplest method for moving object detection, because the background model is simply equal to the previous frame. After performing a binarization process with a predefined threshold with the differencing method, we can find the target contour and the target blob is obtained through the contour filling process. However, sometimes the blob contains too many background pixels when the target is moving very fast, or the blob may lose part of the target information when the target is moving slowly. It is impossible to obtain pure foreground pixels when using the frame differences as the background model, but by using the following method, we can remove the background pixels and retrieve more foreground pixels, on the condition that the color of the foreground is not similar to the pixels of the nearby background. By segmenting the foreground and background in a rectanglular area separately, we can label and cluster the image in the rectangle area again to obtain a more accurate foreground blob.

Fig. 5 Foreground detection process (**a**–**d**)

Figure 5 shows the foreground detection process in frame differencing method, where (a) and (b) show the target detection result using the frame differencing method and the blob filling result, respectively. In Fig. 5c, we can see the left leg of the target is lost. After labeling and clustering process, we can retrieve the left leg (see Fig. 5d).

2.3 Feature Selection and Segmentation

Feature selection is very important in tracking applications. Good features will result in excellent performance, whereas poor features will restrict the ability to distinguish the target from the background in the feature space. In general, the most desirable property of a visual feature is its uniqueness in the environment. Feature selection is closely related to object representation, in which the object edge or shape feature is used as the feature for contour-based representation and color is used as a feature for histogram-based appearance representations. Some tracking algorithms use a combination of these features. In this paper, we use color-spatial information for the feature selection. The apparent color of an object is influenced primarily by the spectral power distribution of the illuminant and the surface reflectance properties of the object. The choice of color space also influence the tracking process. Usually, digital images are represented in the RGB (red, green, blue) color space, but the RGB space is not a perceptually uniform color space because the differences between the colors do not correspond to the color differences perceived by humans. Additionally, the RGB dimensions are highly correlated. HSV (Hue, Saturation, Value) is an approximately uniform color space, that is more similar to human perception and we therefore select this color space for this research. Using the color information is not sufficient, but if we consider combining the color spatial distribution information, then the selected features will become more discriminative.

The main task is to segment the image using this feature. We choose the SMOG method to model the appearance of an object and define the Mahalanobis distance and similarity measure [7]. We then employ the K-means algorithm followed by a standard EM algorithm to cluster the pixels. The difference of our approach is that we do not cluster and track the whole region in rectangle, but only the moving target in the rectangle, as described in the previous section. Figure 6 shows the clustering and tracking results for the whole region in the rectangle, and Fig. 7 shows the clustering and tracking results of the moving target.

If we employ standard method to cluster and track the whole region in the rectangle, it may track properly but requires more particles and computation time.

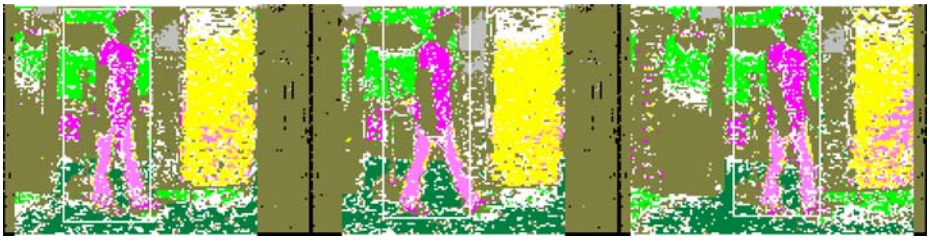

Fig. 6 Clustering and tracking results on the whole region in rectangle

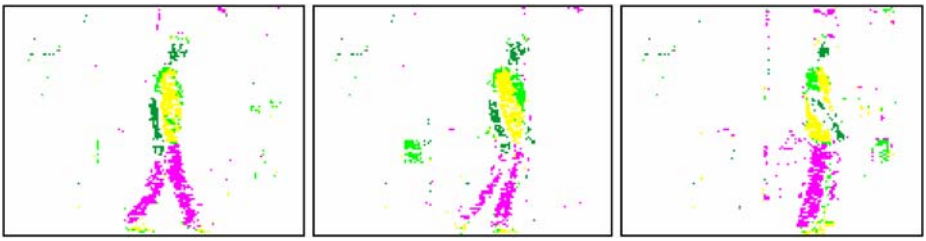

Fig. 7 Clustering and tracking results on moving target

It is because the whole region in the rectangle contains many background pixels. When we choose a new particle at any place in the image, the similarity coefficient is likely high. Thus more particles are required to find good candidate particles from the complicated background. We test the standard method to track a target in real case and the frame rate can reach 10 frames per second at resolution 160*120.

On the other hand, if we cluster and track the moving target only, few particles and less time are needed. The frame rate can reach 15 frames when we track a target employing our method. To save computation time, therefore, the robot clusters and tracks the moving target only.

3 Video Surveillance

3.1 Tracking Using Particle Filter

We propose a tracking strategy that always keeps the target in the scene. Here we do not want and no need always keep the target in the exact center of the scene, because this needs the camera moves frequently and thus it is hard to obtain accurate speed of the target. See Fig. 8, when the target center is in the place between the left edge and right edge, the camera and the mobile platform both remain static. When the target

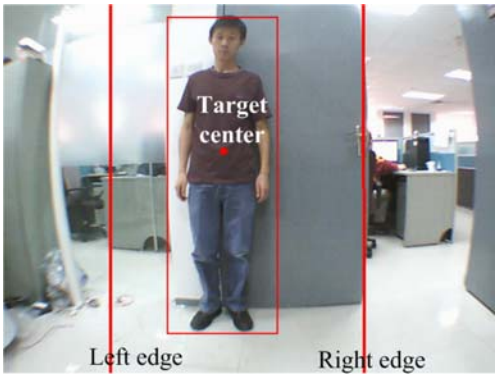

Fig. 8 The left edge and right edge for the tracking strategy

moves and the target center reaches the left edge or right edge, the robot moves to keep the target in the center of the scene according to the predicted results. When the camera moves, particle filtering algorithm is employed to perform the tracking because it can overcomes the difficulty of background changes.

Sequential Monte Carlo techniques, which are also known as particle filtering and condensation algorithms [8–10], have been widely applied in visual tracking in recent years. The general concept is that if the integrals required for a Bayesian recursive filter cannot be solved analytically, then the posterior probabilities can be represented by a set of randomly chosen weighted samples. The posterior state distribution $p(x_k|Z_k)$ needs to be calculated at each time step. In the Bayesian sequential estimation, the filter distribution can be computed by the following two-step recursion.

Prediction step:

$$\mathrm{p}(x_k|Z_{k-1}) = \int p(x_k|x_{k-1})p(x_{k-1}|Z_{k-1})dx_{k-1} \tag{1}$$

Filtering step:

$$\mathrm{p}(x_k|Z_k) \propto p(z_k|x_k)p(x_k|Z_{k-1}) \tag{2}$$

Based on a weighted set of samples $\left\{x_{k-1}^{(i)}, \omega_k^{(i)}\right\}_{i=1}^{N}$ approximately distributed according to $p\left(x_{k-1}|Z_{k-1}\right)$, we draw particles from a suitable proposal distribution, i.e., $x_k^{(i)} \sim q_p(x_k|x_{k-1}^{(i)}, z_k), i = 1, \ldots, N$. The weights of new particles become:

$$\mathrm{w}_k^{(i)} \propto w_{k-1}^{(i)} \frac{p\left(z_k|x_k^{(i)}\right) p\left(x_k^{(i)}|x_{k-1}^{(i)}\right)}{q_p\left(x_k|x_{k-1}^{(i)}, z_k\right)} \tag{3}$$

The observation likelihood function $p(z_k|x_k)$ is important because it determines the weights of the particles and thereby significantly influences the tracking performance. Our observation likelihood function is defined by the SMOG method, which combines spatial layout and color information, as explained below.

Suppose that the template is segmented into k clusters, as described in previous section. For each cluster, calculate its histograms, vertically and horizontally; and make a new histogram of $\tilde{\mathbf{q}}^i$, by concatenating the vertical and horizontal histograms and then being normalized.

As for as a particle is concerned, first classify the pixels into the k clusters above of the template. Then, calculate its normalized histograms of $\{\mathbf{q}_t^i(x_t)\}$, as done with the template.

As in [20], we employ the following likelihood function for $p(z_k|x_k)$:

$$p(z_k|x_k) \propto \prod_i^k \exp -\lambda D^2\left[\tilde{\mathbf{q}}^i, \mathbf{q}_t^i(x_t)\right]$$

where λ is fixed as 20 [20], and D is the Bhattacharyya similarity coefficient between two normalized histograms $\tilde{\mathbf{q}}^i$ and $\mathbf{q}_t^i(x_t)$:

$$D[\tilde{\mathbf{q}}^i, \mathbf{q}_t^i(x_t)] = \left[1 - \sum_k \sqrt{\tilde{\mathbf{q}}^i(k)\mathbf{q}_t^i(k; x_t)}\right]^{\frac{1}{2}}$$

The steps in the particle sample and updating process are as follows.

Step 1 Initialization: draw a set of particles uniformly.
Step 2 (a) Sample the position of the particle from the proposal distribution. b) Find the feature of the moving object. (c) Update the weight of the particles. (d) Normalize the weight of the particles.
Step 3 Output the mean position of the particles that can be used to approximate the posterior distribution.

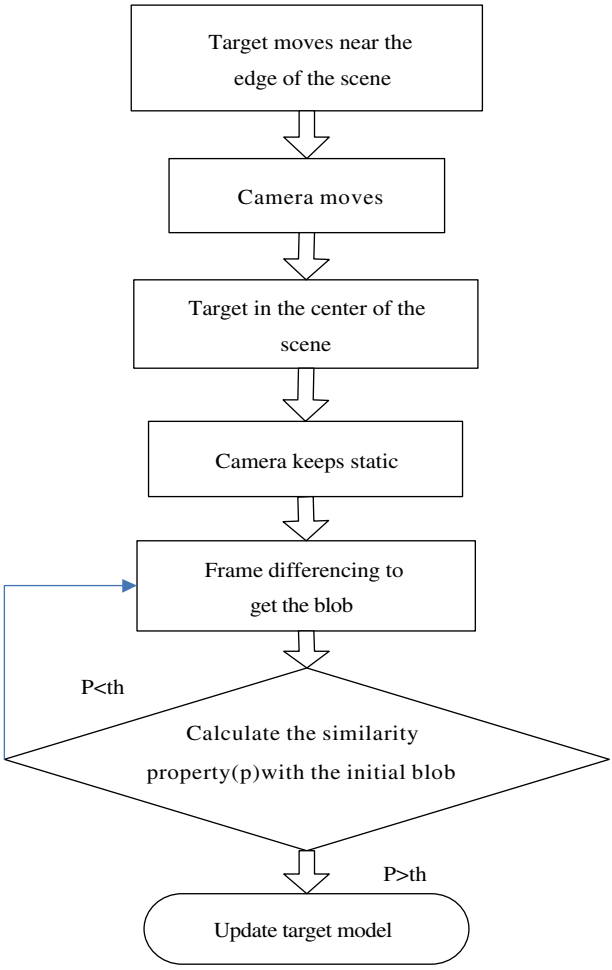

Fig. 9 Target model update process

Fig. 10 Tracking results without update

Step 4 Resample the particles with probability to obtain independent and identically distributed random particles.
Step 5 Go to the sampling step.

Computationally, the crux of the PF algorithm lies in the calculation of the likelihood.

3.2 Target Model Update

The target model obtained in the initialization process cannot be used for the whole tracking process due to changes in lighting, background environment, and target

Fig. 11 Tracking results employing our update process

gestures. We therefore need to update the tracking target model in time. But if we update the target model at an improper time, such as when the camera is moving and the image is not clear, then the tracking will fail. Figure 9 shows a new target model updating process. In our camera control strategy, the camera remains static when the target is in the center of the camera view. When the camera is static, the frame differencing method is employed to obtain the target blob, and the similarity between the current blob and the initial blob is calculated. If the similarity property is larger than a given threshold, then we update the target model, otherwise, we move on to the frame differencing step. How to choose the threshold is an interesting problem. If the threshold is very large, the similarity coefficient may be easily lower than the threshold. This will cause continually updating and consume much computation time. If the threshold is very low, wrong detection will happen. To balance the computation time and tracking results, we choose the threshold as 0.6 according to many experiments in the real environment.

Figure 10 shows failed tracking results without updating in the lighting changes environment. Figure 11 shows robust tracking results employing our update process in the lighting changes environment.

3.3 Experiment Results on Abnormal Behavior Detection

Nowadays, abnormal behavior detection is a popular research topic, and many studies have presented methods to detect abnormal behavior. Our surveillance robot mainly focuses on the household environment, so it is important to detect abnormal behaviors such as people falling down and people running.

It is not easy to track the whole body of a person because of the large range of possible body gestures, which can lead to false tracking. To solve this problem, we propose a method that only tracks the upper body of a person (Fig. 12), which does not vary much with gesture changes. We take the upper half rectangle as the upper body of a target. It may contain some part of legs or lost some part of the upper body. We can obtain pure upper body by using the clustering method mentioned above. Based on this robust tracking system, we can obtain the speed of the target, the height and width of the target. Through the speed of the upper body and the thresholds selected by experiments, running movement can be successfully detected. Also, based on the height and width of the target, we can detect falling down movement through shape analysis.

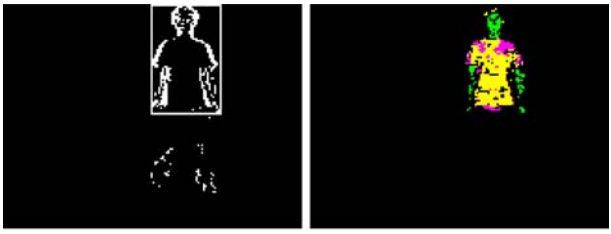

Fig. 12 Clustering results on the upper body

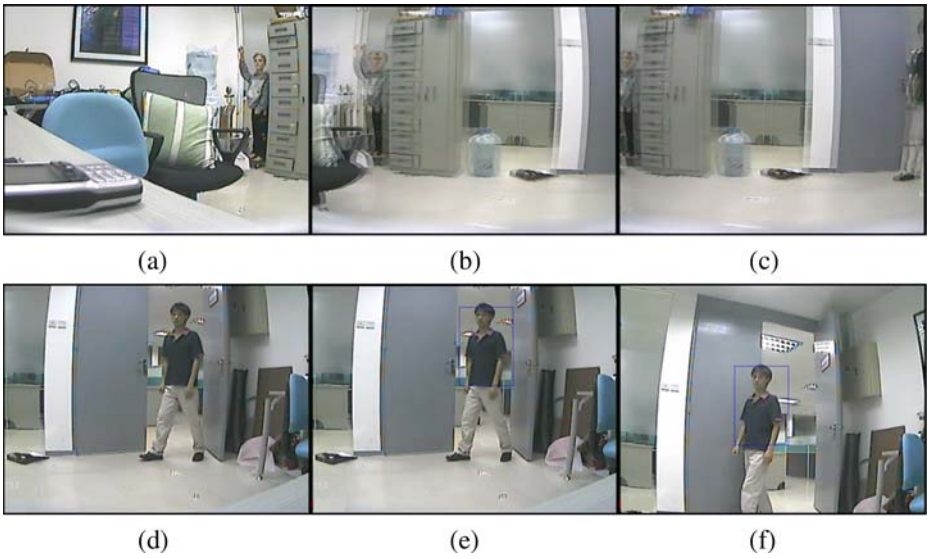

Fig. 13 Initialization and tracking results (**a**–**f**)

Figure 13a–d shows the robot moving to the proper direction to detect the target on receiving a direction signal from the passive acoustic location device. Figure 13a and b are not clear because the camera is moving very quickly. Whereas in Fig. 13e and f, the robot tracks the target and keeps it in the center of the camera view.

Figure 14 shows the detection of the abnormal behavior such as people falling down, bending down, and running in the household environment based on the tracking results.

Fig. 14 Abnormal behavior detection results

4 Abnormal Audio Information Detection

Compared with video surveillance, audio surveillance do not need to "watch" the scene directly. The effectiveness of audio surveillance is not influenced by the occlusions which may cause the failure of the video surveillance system. Especially, in house or storehouse, some areas may be occluded by moving objects or static objects. Also, the robot and people may not in the same room if there are several rooms in a house.

We propose a supervised learning based approach to audio surveillance in household environment [22]. Figure 15 shows the training framework of our approach. Firstly, we collected a sound-effect dataset (See Table 1) which includes many sound effects collected from household environment. Secondly, we manually labeled these sound effects as abnormal samples (e.g. screaming, gun shooting, glass breaking sound and banging sound) or normal samples (e.g. speech, footstep, shower sound and phone ringing). Thirdly, MFCC feature was extracted from a 1.0 s waveform of each sound effect sample. Finally, we trained a classifier using support vector machine. For detecting, when a new 1.0 s waveform was received, MFCC feature was extracted from the waveform; then the classifier was employed to determine whether this sound sample is normal or abnormal.

4.1 MFCC Feature Extraction

To discriminate normal sounds from abnormal sounds, a meaningful acoustic feature must be extracted from the waveform of the sound.

Many audio feature extraction methods have been proposed for different audio classification applications. For speech and music discrimination tasks, the spectral centroid, zero-crossing rate, percentage of "low-energy" frames, and spectral "flux" methods [11] have been used. Spectral centroid represents the "balancing point" of the spectral power distribution. Zero-crossing rate measures the dominant frequency of a signal. The percentage of "Low-Energy" frames describes the skewness of the energy distribution. Spectral "flux" measures the rate of change of the sound. For automatic genre classification, timbral features, rhythmic features, and pitch feature [12], which describe the timbral, rhythmic and pitch characteristics of the music, respectively, have been proposed.

In our approach, the MFCC feature is employed to represent audio signals. The idea of the MFCC feature is motivated by perceptual or computational considerations. As the feature captures some of the crucial properties used in human hearing, it is ideal for general audio discrimination. The MFCC feature has been successfully

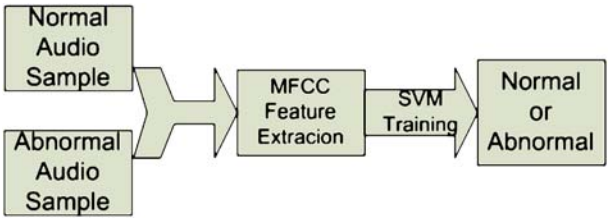

Fig. 15 Training framework

Table 1 The sound effects dataset

Normal sound effects	Abnormal sound effects
Normal speech	Gun shot
Boiling water	Glass breaking
Dish-washer	Screaming
Door closing	Banging
Door opening	Explosions
Door creaking	Crying
Door locking	Kicking a door
Fan	Groaning
Hair dryer	
Phone ringing	
Pouring liquid	
Shower	
…	…

applied to speech recognition [13], music modeling [14], and audio information retrieval [15], and more recently, has been used in audio surveillance [16].

The steps to extract MFCC feature from the waveform are as follows:

Step 1 Normalize the waveform to the range [−1.0, 1.0] and window the waveform with a hamming window;

Step 2 Divide the waveform into N frames, i.e. $\frac{1000}{N}$ ms for each frame;

Step 3 Take the Fast Fourier Transform (FFT) of each frame for getting the frequency information of each frame;

Step 4 Convert the FFT data into filter bank outputs. Since the lower frequencies are perceptually more important than the higher frequencies, the 13 filters allocated below 1000 Hz are linearly spaced (133.33 Hz between center frequencies) and the 27 filters allocated above 1000 Hz are spaced logarith-

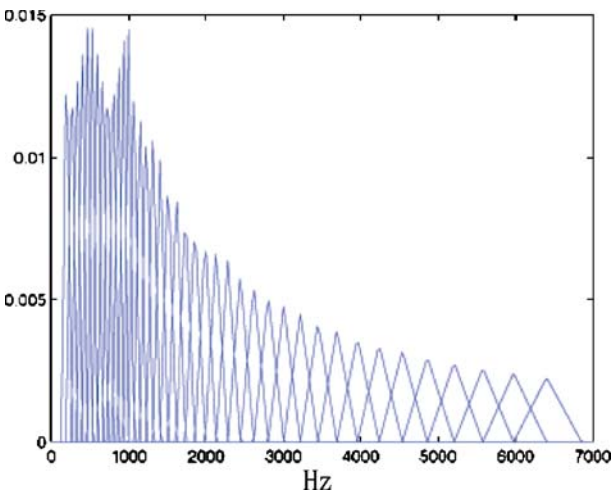

Fig. 16 Frequency response of the triangular filters

mically (separated by a factor of 1.0711703 in frequency). Figure 16 shows the frequency response of the triangular filters;

Step 5 Since the perceived loudness of a signal has been found to be approximately logarithmic, we take the log of the filter bank outputs;

Step 6 Take the cosine transform to reduce dimensionality. Since the filter bank outputs calculated for each frame are highly correlated, we take the cosine transform which approximates principal components analysis to decorrelate the outputs and reduce dimensionality. 13 (or so) cepstral features are obtained for each frame by this transform. If we divide the waveform into 10 frames, the total dimensionality of the MFCC feature for the 1.0 s waveform is 130.

4.2 Support Vector Machine

After extracting MFCC features from the waveform, we employed a classifier trained by support vector machine (SVM) to determine whether this sound is normal or not [17–19].

Our goal is to separate sounds into two classes, normal and abnormal, according to a group of features. There are many types of neural networks that can be used for such a binary classification problem, such as Support Vector Machines, Radial Basis Function Networks, Nearest Neighbor Algorithm, Fisher Linear Discriminant and so on. SVM is chosen as audio classifiers because it has stronger theory-interpretation and better generalization than mentioned neural networks. Compared to other neural network classifiers, SVM has three distinct characteristics. First, it estimates a classification using a set of linear functions that are defined in a high-dimensional feature space. Second, SVM carries out the classify estimation by risk minimization, where risk is measured using Vapnik's ε-insensitive loss function. Third, it implements the Structural Risk Minimization (SRM) principle, which minimizes the risk function that consists of the empirical error and a regularized term.

4.3 Experimental Results on Abnormal Sound Detection

To evaluate our approach, we collected 169 sound effects samples from internet (http://www.grsites.com) including 128 normal samples and 41 abnormal samples (Most of them were collected in household environment).

For each sample, the first 1.0 s waveform was used for training and testing. The rigorous jack-knifing cross-validation procedure, which reduces the risk of overstating the results, was used to estimate the classification performance. For a dataset with M samples, we chose $M - 1$ samples to train the classifier, then tested

Table 2 Accuracy rates (%) using the MFCC feature trained with 20, 50, 100, and 500 ms frame sizes

Frame size	Accuracy
20 ms	86.98
50 ms	86.39
100 ms	88.17
500 ms	79.88

Table 3 Accuracy rates (%) using the MFCC and PCA features trained with 20, 50, 100, and 500 ms frame sizes

Frame size	Accuracy
20 ms	84.62
50 ms	78.70
100 ms	82.84
500 ms	76.33

the performance using the left sample. This procedure was then repeated for M times. The final estimation result was obtained by averaging the M accuracy rates. To train a classifier by using SVM, we apply polynomial kernel where the kernel parameter $d = 2$, and the adjusting parameter C in the loss function is set to 1.

Table 2 shows the accuracy rates using the MFCC feature trained with different frame sizes. Table 3 shows the accuracy rates using MFCC and PCA features trained with different frame sizes. PCA is employed to reduce the data dimension. For example, PCA reduces the data dimension from 637 to 300 when the frame size is 20 ms and from 247 to 114 when the frame size is 50 ms. Comparing the results in Tables 2 and 3, we found that it is unnecessary to use PCA. We obtained the best accuracy rate of 88.17% with a 100 ms frame size. The data dimension of this frame size is 117 which does not need reducing.

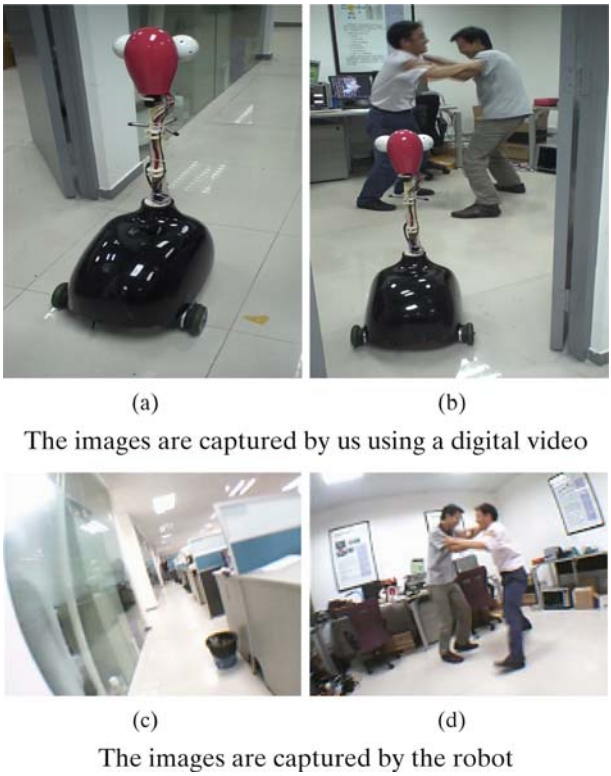

Fig. 17 The process of abnormal behavior detection utilizing video and audio information: **a** The initial state of the robot; **b** The robot turning to abnormal direction; **c** The initial image captured by the robot; **d** The image captured by the robot after turning to abnormal direction

5 Experimental Results Utilizing Video and Audio Information

We test our abnormal sound detection system in real working environment. The sound consists of normal speeches, door opening, pressing keyboard, etc., which are all normal sound. The system gives 8 false alarms in one hour. On the other hand, a sound box is used to play abnormal sound like gun shooting, crying, etc. The accurate rate is 83% among 100 abnormal sound samples. The accuracy is lower comparing with previous experiments for the noise in real environment.

To lower the false alarms by abnormal sound detection system, the passive acoustic location device directs the robot to the scene where abnormal events occur and the robot can employ its camera to further confirm the occurrence of the events. In Fig. 17, we can see the process of abnormal behavior detection utilizing video and audio information when 2 persons are fighting in other room. First, the abnormal sound detection system detects abnormal sound (fighting and shouting). On the same time, the passive acoustic location device obtains the direction. Then the robot turns to the abnormal direction and captures images to check if abnormal behavior occurs. Here we can employ optical flow method to detect fighting, which we presented in previous paper [21].

There are some cases that our system can not detect abnormal behaviors. For example, a person is already falling down before the robot turning to the abnormal direction. To solve this problem, the robot will send the image captured by the robot to the mobile phone of master (Fig. 18).

Fig. 18 The robot sends the image to the mobile phone of master

6 Conclusions

In this paper, we described a household surveillance robot that can detect abnormal events combining video and audio surveillance. Our robot first detects a moving target by sound localization, and then tracks it across a large field of vision using a pan/tilt camera platform. It can detect abnormal behavior in a cluttered environment, such as a person suddenly running or falling down on the floor, and can also detect abnormal audio information and employ its camera to further check the event.

There are three main contributions in our research: a) an innovative strategy for the detection of abnormal events by utilizing video and audio information; b) an initialization process that employs a passive acoustic location device to help the robot detect moving targets; and c) an update mechanism to update the target model regularly.

However, there remains a lot of work to be completed. In future, the abnormal audio information need be realized in hardware like arm systems. To promote the abnormal events detection accurate rate, the fusion of video and audio information need more investigation.

Acknowledgements The authors would like to thank Mr. Y.J. Liang, Mr. Deng Lei, Mr. Qin Jianzhao, and Mr. Fang Zhou for their valuable contribution to this project. The authors would also wish to acknowledgment Mr. Shi Xi with the help of SVM classification process.

The work described in this paper is partially supported by the grant from the Ministry of Science and Technology, The Peoples Republic of China (International Science and Technology Cooperation Projects 2006DFB73360), and by the National Basic Research Program of China (No.2007cb311005).

References

1. Haritaoglu, I., Harwood, D., Davis, L.S.: W4: real-time surveillance of people and their activities. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 8 (2000)
2. Olson, T., Brill, F.: Moving object detection and event recognition algorithms for smart cameras. In: Proc. DARPA Image Understanding Workshop, pp. 159–175, New Orleans, May 1997
3. Zhao, T., Nevatia R.: Tracking multiple humans in complex situations. IEEE Trans. Pattern Anal. Mach. Intell. **26**, 9 (2004)
4. Radhakrishnan, R., Divakaran, A.: Systematic acquisition of audio classes for elevator surveillance. In: Proc. of SPIE, pp. 64–71. Austin, 24–26 May 2005
5. Luo, R.C., Su, K.L.: A multiagent multisensor based real-time sensory control system for intelligent security robot. In: Proceedings of International Conference on Robotics and Automation, Taiwan, 14–19 September 2003
6. Massios, N., Voorbraak, F.: Hierarchical decision-theoretic planning for autonomous robotic surveillance Massios. In: Advanced Mobile Robots, 1999 Third European Workshop, pp. 219–226. Zurich, 6–8 September 1999
7. Wang, H., Suter, D., Schindler, K., Shen, C.: TAdaptive object tracking based on an effective appearance filter. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 9 (2007)
8. Khan, Z., Balch, T., Dellaert, F.: MCMC-based particle filtering for tracking a variable number of interacting targets. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 11 (2005)
9. Carpenter, J., Clifford, P., Fernhead, P.: An improved particle filter for non-linear problems. Technical Report, Dept. Statistics, Univ. of Oxford (1997)
10. Arulampalam, M.S., Maskell, S., Gordon N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process. **50**, 2 (2002)
11. Scheirer, E., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In: Proc. of International Conference on Acoustics, Speech, and Signal Processing, pp. 1331–1334. Munich, 21–24 April 1997

12. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Trans. Speech Audio Process. **10**(5), 293–302 (2002)
13. Holmes, J.N., Holmes, W.J.: Speech Synthesis and Recognition, 2nd edn. Taylor & Francis CRC, London (2001)
14. Logan, B.T.: Mel frequency cepstral coefficients for music modeling. In: Proceedings of the First International Symposium on Music Information Retrieval, Bloomington, 15–17 October 2001
15. Foote, J.: An overview of audio information retrieval. Multimedia Syst. **7**(1), 2–10 (1999)
16. Radhakrishnan, R., Divakaran, A., Smaragdis, P.: Audio analysis for surveillance applications. In: IEEE Workshop on Application of Signal Processing to Audio and Acoustics, pp. 158–161. New Paltz, 16–19 October 2005
17. Cristianini, N., Shawe-Taylor, J.: A Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge (2000)
18. Bernhard, S., Burges, C.J.C., Smola, A.: Advanced in Kernel Methods Support Vector Learning. MIT, Cambridge (1998)
19. Ou, Y., Wu, X.Y., Qian, H.H., Xu, Y.S.: A real time race classification system. In: Information Acquisition, 2005 IEEE International Conference. Hong Kong, 27 June–3 July 2005
20. Perez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based probabilistic tracking. In: European Conference on Computer Vision, pp. 661–675. Copenhagen, 27 May–2 June 2002
21. Ou, Y.S., Qian, H.H., Wu, X.Y., Xu, Y.S.: Real-time surveillance based on human behavior analysis. Int. J. Inf. Acquis. **2**(4), 353–365 (December 2005)
22. Wu, X.Y., Qin, J.Z., Cheng, J., Xu, Y.S.: Detecting audio abnormal information. In: The 13th International Conference on Advanced Robotics, pp. 550–554. Jeju, 21–24 August 2007