Advances in Optics and Computer Vision

Journal homepage: http://www.aocv.ir

AOCV

Detection of abnormal human activities in video using combining classifiers

Sarvenaz Sadeghi-ivrigh

Department of Computer, South Tehran Branch, Islamic Azad University, Tehran, Iran
**Corresponding Author, Email: Sarvenaz.sadeghi@gmail.com*

ARTICLE INFO

Article history:
- *Received: 2017-11-9*
- *Revised: 2017-12-18*
- *Accepted: 2017-12-31*
- *Available online: 2018-1-5*

Keywords:
Abnormal activity detection
Behavior detection
Human detection
Integrated video descriptor

ABSTRACT

Control of public places such as stadiums, banks etc. by CCTV cameras require so many manpower. These systems are vulnerable to error because of fatigue or human error. So, we must develop a system that be able to detect normal and abnormal activities besides aware the security forces about the situation to provide better protection. In this research we introduce an integrated descriptor where after cutting videos, we perform Fourier transform to map the movement information to frequency spectrum/domain and then extract information. We apply Gabor filter for produced frequency spectrum and extract features. Then, we have dimension reduction with JMI feature selection and training with SVM and finally, behavior detection. Non-isolated background from foreground causes that our system works like a referee in sports games for activity detection when we need to foreground and background information simultaneously. With feature selection step in dimension reduction and hierarchical SVM, we show the superiority of this method according to accuracy and speed criterion compared to other methods.

1. Introduction

Control of public places such as stadiums, banks etc. by CCTV cameras require so many manpower. These systems are vulnerable to error because of fatigue or human error. So, we must develop a system that be able to detect normal and abnormal activities besides aware the security forces about the situation to provide better protection. This system is a non-smart supervised system. So, in recent years, increased attention has been taken to design a system that can identify normal and abnormal activities and aware to the security forces. In previous studies, all human activity recognition systems follow these 4 steps. You can see this in figure1.

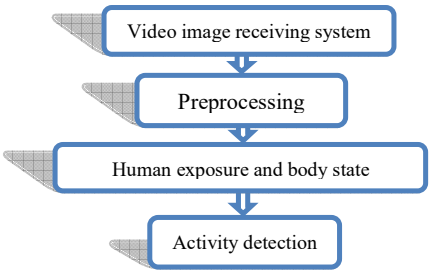

Figure 1. Common human activity detection system

Several studies in the literature compete in third and fourth stages for human activity detection. In the proposed method in the first stage, first of all, we cut the received 3-D films collections and then extract foreground and background features simultaneously instead of processing. Then, to facilitate the movement information extraction, we map foreground and background information to frequency domain. This is an answer to judgment supervision systems for sports games that needs to foreground and background information simultaneously. Instead of the third stage of previous studies that was human exposure, we reduced dimensions with JMI[1] feature selection and spatial transform. In the fourth stage, using hierarchical SVM[2] , we increase accuracy and speed of activity detection. In the second part of this article, we present related research and in the third part we review the proposed method of this article. In the fourth part, we will have proposed method and in the last part, we present conclusion an ẋd future work.

[1]Joint Mutual Information
[2]Support Vector Machine

2 .Empirical background

In the last few years, increased attention has been paid to supervision systems with the aim of increase the interaction between human and machine. The proposed activity detection system follows 4 steps. Some kinds of actions have been taken in preprocessing phase to reduce under desired domain reduction using background subtraction and luminous flux. A lot of background modeling methods has been developed by Krystany et.al in 2010 and Elhabian in 2008 [1]. Find a displacement value between consecutive frames from an image is called luminous flux. The first method for variable luminous flux calculation in an image sequence had been introduced by Horn and his colleagues in 1981 [2]. The most sensitive and the most important part of the system is human exposure and body state estimation. This part is divided into two partitions. The most common methods are component-based methods and single identification window analysis [3]. In the last stage, both activity and action detection methods can be classified in two single-layered methods. Single layer method is suitable for action detection and hierarchical method is suitable for activity detection [4]. In 2010, Foresti et.al presented an article about supervised vector event detection in video sequences using single-class Supporting Vector Machine [5]. In 2014, Chen and his colleagues [6] introduced a new algorithm for human activity detection in video on the basis of network with SVM [7]. Also Kang et.al presented abnormal behavior detection algorithm using combined agents in crowded scenes in 2014. In 2016, Yeo et.al [8] introduced an unsupervised activity detection algorithm on the basis of Markov model. In 2015, Ommer and his colleagues [9] presented an article about video temporal-spatial analysis for disorder detection using graphical model and video analysis. All of the previous methods were based on movement and was needed background estimation and foreground exposure or worked on the basis of luminous flux caused by human movement. In the both states, all of the preprocessing steps, foreground exposure and finally, perform a suitable classifier algorithm are time and memory consuming. Also, in some activities such as judgment in sports games, we need to scene information and leg or hand movement information simultaneously. Using previous methods such as background detection or body states extraction we couldn't reach to this important information.

3 .Proposed method

Previous studies were needed to estimate background and foreground exposure. In this state, total preprocessing step, foreground exposure and finally, perform a suitable classification algorithm were time and memory consuming. were time and memory consuming. So, we present a method that needs to different body states extraction in different frames and non-isolated background from foreground information. This method can improve system performance and reduce the time required. You can see general framework of proposed method in figure2.

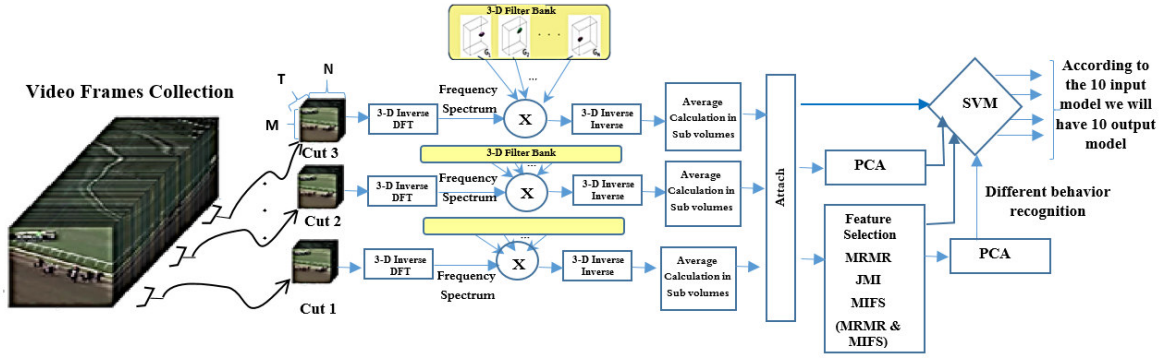

Figure 2. General Frame Work of Proposed model

The non-isolated foreground from background information property during feature extraction is suitable for sports games judgment too. This need and some above mentioned undesirable conditions encouraged us to design an integrated video descriptor that extracts movement and scene information simultaneously. Then with the help of feature selection we reduced dimensions desirably. In this method accuracy didn't reduced and in this stage, we were also faced with increasing accuracy. This is differentiation point of proposed method. First, we cut 3D video frames set. We use 3 cuts in experiment. The calculation reduction and more than 3 cut selection didn't affect performance improvement. We map video films cuts to frequency domain because activity detection in frequency domain is easier. Here, we assume that camera is fixed so background information is equal in all video frames. Although, convolution in primary domain (spatial) is equal to multiply in frequency domain and because of background information, we can separate frequency domain from background information easily. To map video frames set to frequency domain, we used discrete Fourier transform. F (x, y, t) 3-D Fourier transform on space and time calculates as follows:

$$F(u,v,w) = \frac{1}{MNT} \times \sum_{x=0}^{M-1}\sum_{y=0}^{N-1}\sum_{t=0}^{T-1} f(x,y,t)\, e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}+\frac{wt}{T}\right)} \quad (1)$$

Where M, N, T are width, height and time for video cut respectively. x, y, t are spatial and temporal locations of every point in created volume. You can see the frequency spectrum presentation in figure3.

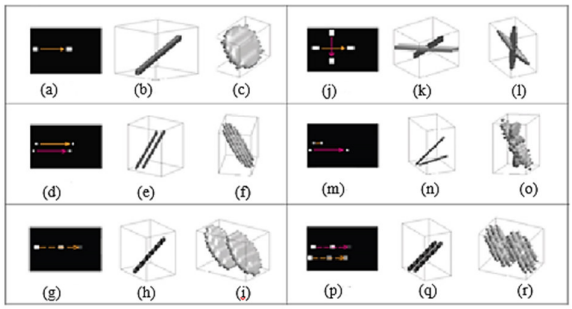

Figure 3. Frequency spectrum presentation

To extract features, we used 3D Gabor filter bank. Gabor filter is modeled structure of human eyes and it is a suitable descriptor for movement. This filter finds edges in different directions and describes its related movements. You can see filter presentation in figure4.

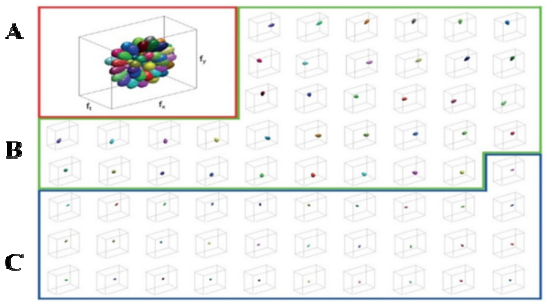

Figure 4. 3-D filter presentation, a) First and second step filters together, b) All filters from first step, c) Second step filters

Transfer function of each 3-D filter is equal to fr0 spatial frequency along specified direction with θ0 polar angles and Φ0 direction in spherical coordinate system and we can say this as follows:

$$G(f_r,\theta,\varphi) = \exp\{-\frac{(f_r-f_{r0})^2}{2\sigma_r^2} - \frac{(\theta-\theta_0)^2}{2\sigma_\theta^2} - \frac{(\varphi-\varphi_0)^2}{2\sigma_\varphi^2}\} \quad (2)$$

σ_Φ, σ_θ, σ_r are radial and angular bandwidth. These parameters specify filter stretch in spatial-temporal frequency domain. You can see frequency spectrum filtration effect in figure5.

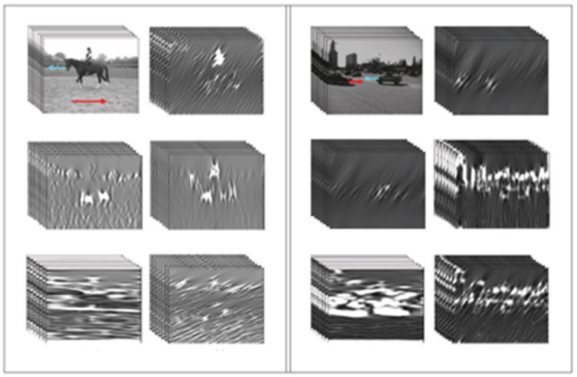

Figure 5. The Effect of Frequency Spectrum filtration

The fourth step of inverse 3D discrete Fourier transform includes extracted features in frequency domain that must return to initial domain.

$$f(x,y,t) = \sum_{u=0}^{M-1}\sum_{v=0}^{N-1}\sum_{w=0}^{T-1} F(u,v,w)\, e^{j2\pi\left(\frac{xu}{M}+\frac{yv}{N}+\frac{tw}{T}\right)} \quad (3)$$

In the fifth step, the average value is calculated from produced sub-volumes in each cuts and send them to the next step (attach) because we need to sub-volumes information in each cuts to train our system with them. In the sixth step, we attach calculated averages from sub volumes to each other. In the next phase, generated feature vector length should be reduced. In this step because the length of generated feature vector reach to 104448, so we should reduce dimension. To reduce generated feature vector dimension, first, we performed PCA[3] but we didn't satisfied. So, we used feature selection such as JMI, MIFS[4], MRMR[5], (MRMR & MIFS) that were the best methods in dimension reduction and they had highest accuracy. Among these methods JMI was the best method. You can see the results in table1. In the last stage, we classified with hierarchical SVM and detected behaviors. We attached a label to each data in the first step of SVM then in the second stage used labeled data probabilities in first stage as new inputs to SVM because we wanted to label the unlabeled data.

[3]Principal Component Analysis
[4]Mutual Information-based Feature Selection
[5]Maximum Relevance Minimum Redundancy

Table 1: The comparison between 4 methods according to dimensions' reduction and max accuracy in both banks

The number of max accuracy in UCF50	Max accuracy in UCF50 with 3 iteration	Dimension reduction in UCF50 bank	The Number of max accuracy in UCSD	Max accuracy in UCSD with 30 iteration	Dimension reduction in UCSD bank	Methodologies
3	0.6873	6681*104448	1	0.8571	70*104448	All Data
1	0.5667	6681*1336	8	0.5750	70*70	All Data+ PCA
2	0.6740	6681*6357	29	1	70*18695	MRMR
2	0.6650	6681*6060	29	1	70*18696	MIFS
3	0.6890	6681*6884	4	1	70*18150	JMI
2	0.6326	6681*2732	29	1	70*16530	MRMR& MIFS
1	0.5350	6681*1336	8	0.5750	70*70	MRMR+PCA
2	0.5170	6681*1336	8	0.5750	70*70	MIFS+ PCA
3, 1	0.5475	6681*1336	8	0.5750	70*70	JMI+ PCA
1	0.4862	6681*1336	8	0.5750	70*70	(MRMR &MIFS) +PCA

4 .Experiment results

In order to evaluate the proposed method, we applied two dataset called UCF50 (htt) and UCSD (htt1). The UCF50 dataset includes web videos, 50 categories of different activities in uncontrolled situation with more than 100 video for each category. Also, UCSD dataset includes pedestrian crossing with or without vehicle. You can see some available video samples in this dataset in figures 6&7.

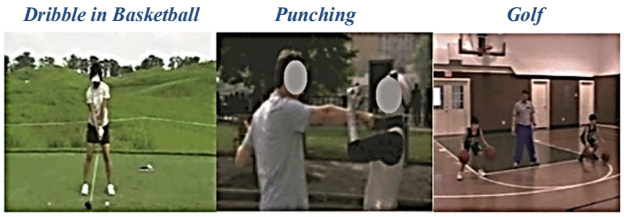

Figure 6. Sample Videos from UCF50 data set

Figure 7. Sample Videos from UCSD data set, a) Normal Activity) b) Abnormal Activity

The feature vector length in our experiment was 104448 that included 68 filters, 512 volumes and 3 video cuts (3*68*512) =104448. You see the 68 applied 3D Gabor filter presentation in figure 4. The results of features selection actions compared in table 1.

a) By comparing the accuracy change diagram (figures 8&9) with10 proposed method for dimension reduction, first of all, we performed PCA and reduced dimension but we didn't satisfy. For this reason, we used feature selection filtering method. You can see these 10 methods includes 3 feature selections, a binary compound and PCA actions in table1. We compared these 10 methods according to accuracy value with 30 iterations in UCSD bank with 70 samples because the number of samples of this bank was lower than UCS50 but the number of iteration was 3. This value was acceptable. You can see this comparison in table 1.

b) According to the second column of table, the main value of dimension reduction is related to PCA applied data. This value is 70*70 in all 5 methods. PCA performed data on total data causes that features dimensions doesn't exceed of samples features. Even though with perform PCA on data, the dimensions reduced more than before but we missed a lot of important information. The comparison between methods in dimension reduction such as MRMR & MIFS, MRMR, MIFS and JMI, we showed that with accuracy value equivalency with value: 1 in the third column, these methods had the best dimension reduction value but in JMI the highest accuracy occurred in fourth replication. This value in the other methods occurred in 29th replication. So, with reasonable and logical looking and according to dimension reduction and max accuracy value and the number of iteration, JMI was the best method. This superiority of JMI in UCF50 bank in fifth column presented in table1 .

c) According to table1, JMI in both UCSD and UCF50 banks had max accuracy value after ALLDATA. Although, JMI had sometimes higher accuracy than ALLDATA, when dimension reduction methods didn't apply and we had initial dimension values. You can see this in table 1. JMI runs presentation with accuracy value=1 in comparison with all data with accuracy value=0.8571 in proposed method in 4folds from 10folds in SVM. This feature is a result of random feature selection filtering that caused we acquired very good features. We present 10 performed methods for dimension reduction for both UCSD and UCF50 banks according to the average iteration number in figure 8 & 9.

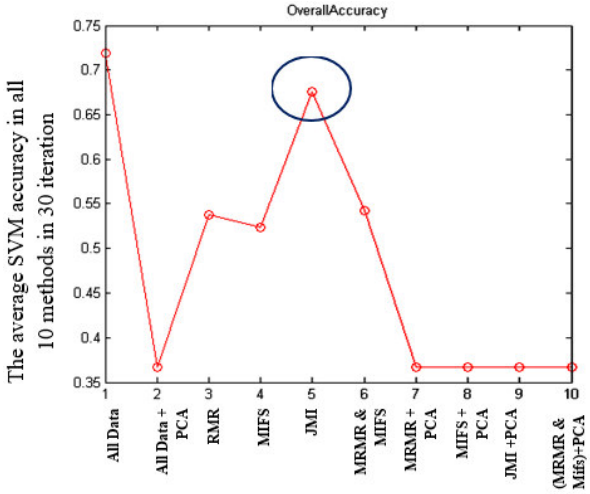

Figure 8. Accuracy average in UCSD with different methods

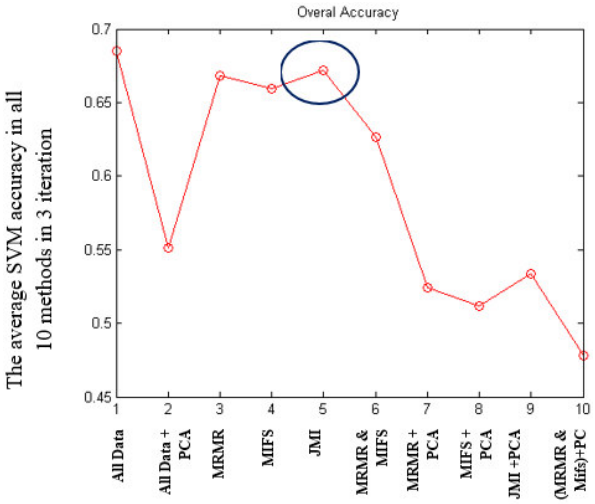

Figure 9. Accuracy average in UCF50 with different methods

To proposed descriptor performance evaluation, we compared STIP[6] [10] and GIST[7] [11] descriptors in UCSD and UCF50 dataset. You can see this in table2. Our descriptor is a suitable method with 69.90 % accuracy in UCF50 with 50 activity category in this dataset. Also, this descriptor has normal and abnormal 70 films samples in UCSD. Our descriptor has 100% accuracy value in this bank. Its highly precise descriptor that is better than 2 other methods.

Table 2. Activity detection results in UCSD and UCF50 banks

Accuracy in UCSD	Accuracy in UCF50	Descriptor
41.3%	42.4%	GIST descriptor
50.5%	52.5%	STIP descriptor
100%	67.90%	3-D integrated descriptor with JMI feature selection

[6]Spatio-Temporal Interest Point
[7]Generalized Integrated Search Tree

5 .Conclusion and future work

Developing a system with protective forces notification without fatigue and confused, bring better security for us. Therefore, in this article, we presented an integrated video descriptor for human activity detection. In this method, we used Fourier transform for video cuts. Then, mapped the movement information to frequency domain and then extracted the movement information. Then, we performed Gabor filter for generated frequency spectrum and extracted features. Then, we had dimension reduction with feature selection, space conversion and trained with SVM and finally, detected behaviors. According to conclusions, among all methods, JMI was the best method in terms of accuracy against dimension reduction. Among other preprocessing actions in data, we can point to data cleansing, feature subset selection, feature selection, sample filtering, sampling, data conversion, discretization, dimension reduction, data aggregation and feature creation. In this research, we used feature selection and dimension reduction. We will propose some other methods for dimension reduction in the future.

In the training phase, we used SVM but it's desirable that present a training course with higher accuracy for classify data. Other training algorithms for classification accuracy incensement includes: decision tree-based method, rule-based method, memory-based reasoning, neural networks, Bayesian theory based methods, supporting vector machine. To classify data, we used combined classifier. This is one of the aims of this research. For this purpose, we used hierarchical SVM in all phases. Here, we can use other classification algorithms such as decision tree, neural networks or rule-based methods after first stage of SVM. The innovation aspect of the proposed method can be noted different feature selection and non-isolated background from foreground information during feature extraction.

Reference

[1] B. Anti'c, and B. Ommer, "Spatio-temporal Video Parsing for Abnormality Detection". arXiv:1502. 06235[cs. CV], 1, Feb 2015.
[2] S. Elhabian, K.M. El-Sayed, S.H. Ahmed, "Moving object detection in spatial domain using background removal techniques-state-of-art". Recent Patents on computer Science, 1, pp. 32-54, 2008.
[3] D. Gavrial, "The visual analysis of human movement: A Survey". Computer Vision and Image Understanding, 73, pp. 82-98, 1999.
[4] B. Horn, and B. Schunck, "Determining optical flow. Artificial Intelligence", 17, pp. 185-204, 1981
[5] S. Hyun Cho, and H. Bong Kang, "Abnormal behavior detection using hybrid agents in crowded scenes". Pattern Recognition Letters, 44, pp. 64-70, 2014.
[6] W. Lin, Y. Chen, J. Wu, H. Wang, B. Sheng, and H. Li, "A new Network-Based Algorithm for Human Activity Recognition in Videos". IEEE Transactions on circuits and Systems for Video Technology, 24(5), May 2014.
[7] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild". IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[8] O. Oliva, A. Torralba, G. Dugue, and J. Herault, "Global semantic classification of scenes using power spectrum templates". Challenge of Image Retrieval, pp. 1-12., 1999.

[9] C. Piciarelli, and G. Foresti, "Surveillance- oriented event detection in video streams". IEEE Intell. Syst, 26(3), pp. 32-41, 2010.

[10] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-termporal features for action recognition". In British Machine Vision Conference, 2009.

[11] D. Yeo, B. Han, and J. Han, "Unsupervised Co-Activity Detection from Multiple Videos Using Absorbing Markov Chain". Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), 2016.

[12]www.svcl.ucsd.edu/projects/anomaly/UCSD_Anomaly_Dataset.tar.gz.

[13] www.crcv.ucf.deu/data/UCF50php.