Journal of Soft Computing Paradigm (JSCP) (2021)
Vol.03/ No.02
Pages: 55-69
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2021.2.001

An Efficient Dimension Reduction based Fusion of CNN and SVM Model for Detection of Abnormal Incident in Video Surveillance

Dr. Rajesh Sharma,
Computer Vision and Robotics, ASTU,
Adama, Nazret.

Dr. Akey Sungheetha,
Data Science, ASTU,
Adama, Nazret.

Abstract- Performing dimensionality reduction in the camera captured images without any loss is remaining as a big challenge in image processing domain. Generally, camera surveillance system is consuming more volume to store video files in the memory. The normally used video stream will not be sufficient for all the sectors. The abnormal conditions should be analyzed carefully for identifying any crime or mistakes in any type of industries, companies, shops, etc. In order to make it comfortable to analyze the video surveillance within a short time period, the storage of abnormal conditions of the video pictures plays a very significant role. Searching unusual events in a day can be incorporated into the existing model, which will be considered as a supreme benefit of the proposed model. The massive video stream is compressed in preprocessing the proposed learning method is the key of our proposed algorithm. The proposed efficient deep learning framework is based on intelligent anomaly detection in video surveillance in a continuous manner and it is used to reduce the time complexity. The dimensionality reduction of the video captured images has been done by preprocessing the learning process. The proposed pre-trained model is used to reduce the dimension of the extracted image features in a sequence of video frames that remain as the valuable and anomalous events in the frame. The selection of special features from each frame of the video and background subtraction process can reduce the dimension in the framework. The proposed method is a combination of CNN and SVM architecture for the detection of abnormal conditions at video surveillance with the help of an image classification procedure. This research article compares various methods such as

ISSN: 2582-2640 (online)
Submitted:6.03.2021
Revised: 30.03.2021
Accepted: 21.04.2021
Published: 10.05.2021

background subtraction (BS), temporal feature extraction (TFE), and single classifier classification methods.

Keywords: *Deep learning, video surveillance*

1. INTRODUCTION

Recently, video surveillance is automated to reduce the manual labor workload. This automation analysis is increasing day by day for event activity tracking and recognition. The activity tracking and recognition are very challenging, when the person is moving from one window of the camera to another [1]. The automated system is very essential in public areas with less manpower. The detection of abnormal events from the public sector is a challenging task with high accuracy [2]. This video surveillance is developed to detect the following abnormal conditions, such as fighting, robbery, accident, chain snatching etc. A huge amount of cameras are deployed worldwide to ensure the public safety [3]. These cameras are struggling to store all the video stream data including the unnecessary data in their memory. Henceforth, there is a need to incorporate automatic monitoring techniques along with the existing system [4]. Due to the limited performance of many monitoring techniques, classification techniques are very essential in the video surveillance problem recently. Figure 1 shows an example of an image frame obtained from the video.

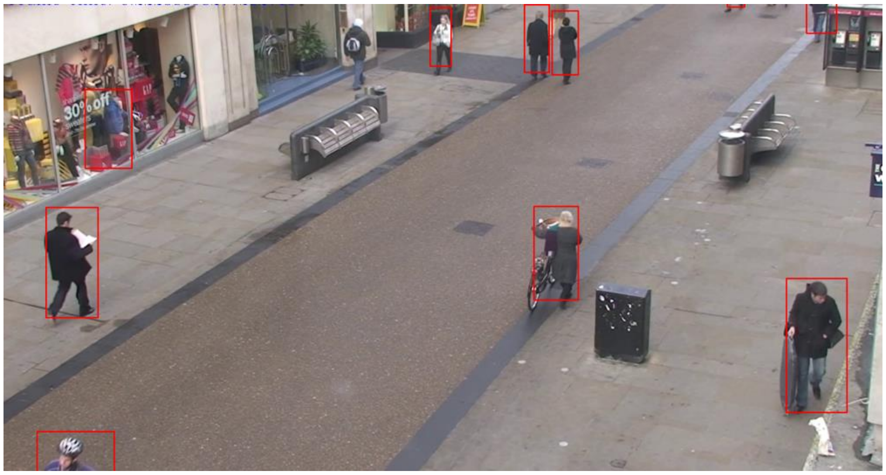

Figure 1 Example of single frame of Video Surveillance

ISSN: 2582-2640 (online)
Submitted:6.03.2021
Revised: 30.03.2021
Accepted: 21.04.2021
Published: 10.05.2021

An efficient automatic computer vision is required to classify the normal and abnormal conditions in the video frame without requiring any manual effort [5]. This automatic method is used to reduce the manual power and monitor public safety. The anomalous activities occur with a variety of normal designs [6]. The visual information is extracted to classify and make a difference between the normal and abnormal events. Violence detection and road accident detection are the real-time complication processes involved in the scenarios [7].

The activity recognition from frames of any videos should be more accurate even though there is a sparse in the frame due to the variation of environment [8]. The algorithm should preprocess and clean the images obtained from the noise sector to find a very clear picture. Sometimes, these variations in the background are unpredictable for many reasons such as rainy, foggy, lighting, occlusion etc. [9]. This variation can be measured with the same action class of different viewpoints. The many action classes have included the transformation of same frames based on the same viewpoints and intra class variation [10]. But the lighting or foggy conditions are sensed by sensors and the variation can be determined with the same action. The scaling function and variation in the video cannot be transformed often. In addition to that, the identical problem in the class can be classified by using the appropriate methods [11].

2. ORGANIZATION OF THE RESEARCH

This research article is organized as follows; section 3 provides a literature review of the paper, section 4 delivers the explanation of the methodology, section 5 investigates the experimental results, which is further followed by the future task and conclusion in section 6.

3. PRELIMINARIES

A lot of differences are observed between the motion-based approach and shape-based approach based on the computation time. The computation time remains very lesser in the motion-based approaches [1]. In more noisy video streams, the tracking of pictures is considered as a very challenging task. To overcome this problem, many methods are developed for performing effective computations. The computation efficiency is improved by using many methods such as spatial-temporal and volume-based methods [11]. Wang et al measure the crowd population by utilizing the point-based feature extraction method. The projection is based

ISSN: 2582-2640 (online)
Submitted:6.03.2021
Revised: 30.03.2021
Accepted: 21.04.2021
Published: 10.05.2021

on interest points and a histogram of the oriented gradient. These approaches are discussed with the meeting of dense set interest points [12]. This temporal dimension is very expensive to compute.

The authors have described the interest point detector for performing motion prediction. They used trajectory-based trackers to detect the motions [10]. The shape-based motion detection is determined by using the extracted features. This procedure can be done by any classification method. Dollar et al investigated the temporal Gabor filter approach with spatial Gaussian filter for motion detection and recognition. Higher recognition accuracy can be achieved even for a sparse set of points [13]. The feature extraction is very important to reduce the dimension of every video frame. The appropriate method like spatial-temporal method is used to extract the features for processing the minimum element of images [12]. Weinland et al have introduced the feature extraction method to extract the features from every frame of video by using motion history volume. This technique detects the features with more sparse and moderate accuracy [14].

Mehran et al examined the social force model, which is used to detect the anomalous events in the road [15]. Kim et al describe Markov random field will detect the local activities from their camera and further categorizes the anomalies activities [16]. Li et al have introduced the detector for identifying anomaly activities in the crowd by using a mixture of dynamic texture models. This method has surveyed well in sparse reconstruction obtained from the camera with outliers in pattern modeling [17]. Cong et al investigated a sparse reconstruction procedure with many abnormal conditions in the view sectors, which contains the sequence of patch images with spatiotemporal features obtained from camera surveillance. These extracted features are spotted and analyzed the comprehensive abnormal incidences [18]. The supervised model approach can be used for the whole dynamic coded program for finding the difference of normal and abnormalities present in the video sequence. This approach learns to detect the abnormal conditions from the surveillance [19]. Lu et al investigated a sparse combination learning model for performing efficient analysis with cloud server [20].

Nowadays, deep learning approach is very popular in recognizing the abnormal and normal activities. The convolutional neural network (CNN) is used for successfully performing image classification in any research article. The image base CNN method classifies the action of any human or animal. This CNN algorithm requires a huge amount of samples for training and testing. Since the camera clarity is very high; training the CNN model will consume more time [21]. The deep learning architecture is used for image classification with more effectiveness [22][23]. Support vector machine is combined with the CNN model for performing image classification in order to recognize the images with abnormal or normal conditions. The feature extraction and recognition are very good in this combination other than many other combinations [24]. Besides, this procedure attains a good accuracy after classifiers train them. Despite the hype, this procedure suffers from the longer duration taken for training the datasets. It consumes more time to train and test because of its high dimension. The proposed method has combined the methodology of various deep learning architectures to reduce the timing of training by mitigating the dimension of the images without any loss. This research work aims to propose a novel method to detect the abnormalities for any group or individual activities that are isolated by using the supervised learning method.

4. METHODOLOGIES

4.1 Background Subtraction Method

Gaussian mixture procedure is used to subtract the background from each frame of the video. The pixel of every mount is framed by using a Gaussian distribution mixture (GDM) [25]. The probability of the pixels is written as,

$$P(x_k) = \sum_{i=1}^{N} w_i \eta(x_k : \theta_i)$$

Normal distribution of nth factors are represented by,

$$\eta\left(x; \mu_n, \sum\quad_n\right) = \frac{1}{(2\pi)^{\frac{D}{2}} |\sum_n|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}}(x-\mu_n)^T \sum_n^{-1} (x-\mu_n)$$

Where, T is minimum fraction, μ_n is mean $\sum_n$ is covariance of the nth components.

The minimum prior model probability in the frame is subtracted from the background and it is performed by setting the pixel value in every normal distribution. The Gaussian distribution is updated and computed for marking the foreground pixels in the images and it also removes the unwanted information [26]. Figure 2 shows the overall workflow of the proposed work.

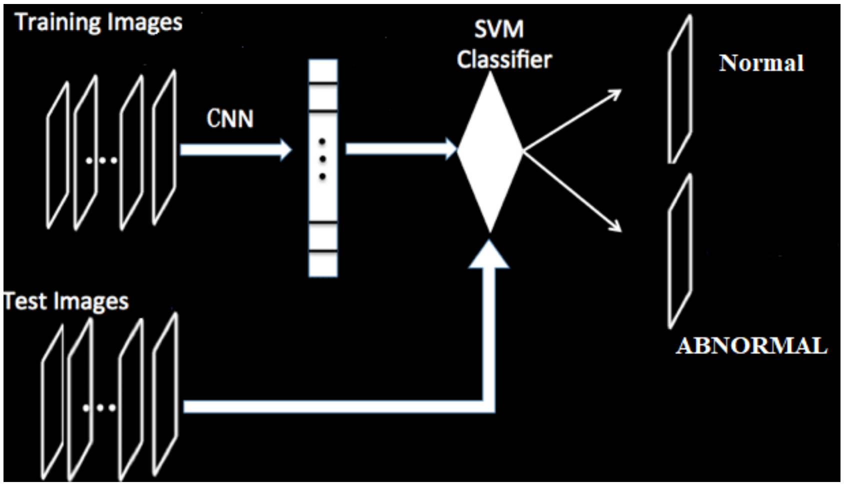

Figure 2 Our proposed framework output

4.2 Spatial feature extraction Approach

The raw input data can be extracted by using a single classifier CNN in order to recognize the image activities/events. This network is used to extract the features from the images in order to recognize the human activities [27]. The networks are constructed with 16 layers along with some static pooling layer, which is shown in figure 4. This depth of the network layer is playing an important role in recognizing the activities but this addition of the layer should be significant. Extra layers will lead to inaccuracy and consume more computation time, which means that it will increase the system complexities [28] [29].

4.3 Proposed frame work

The proposed framework consists of pre-processing and classification sections. The background subtraction method can be incorporated with fused classification frameworks. Figure 3 shows the proposed overall hybrid framework for training and classification arrangement.

ISSN: 2582-2640 (online)
Submitted:6.03.2021
Revised: 30.03.2021
Accepted: 21.04.2021
Published: 10.05.2021

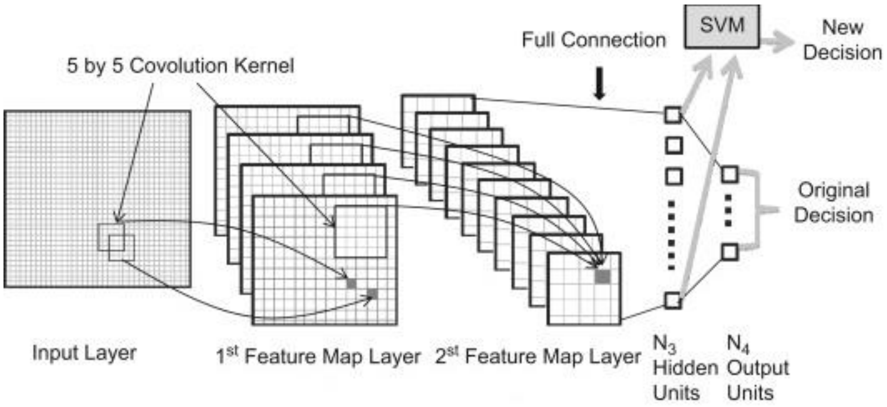

Figure 3 Hybrid Framework of our proposed system

The proposed method has fused two algorithms such as CNN and SVM for training and classification purposes respectively. The output of the proposed hybrid framework contains a softmax activation function to enable the prediction in the output results [30]. Here, it has been integrated with linear SVM classifier to recognize the normal human activities. The SVM includes the state-of-the-art linear classifier techniques in the balanced classes. The proposed algorithm has compressed the dimensional reduction in the sample dataset. This reduction is possible by subtracting the background of each frame present in the model [31]. This combination is providing more dimension reduction in the pictures, which in turn makes the system easier. The feature extraction is computed by using CNN and it is trained for performing accurate classification by using the SVM model.

To classify the activities with a small number of videos are required from the video surveillance. The changes in the video can be classified by the proposed algorithm, which provides good accuracy to predict the correct condition of the human. Here, the supervised learning principle is used for performing feature extraction [32][33][34].

ISSN: 2582-2640 (online)
Submitted:6.03.2021
Revised: 30.03.2021
Accepted: 21.04.2021
Published: 10.05.2021

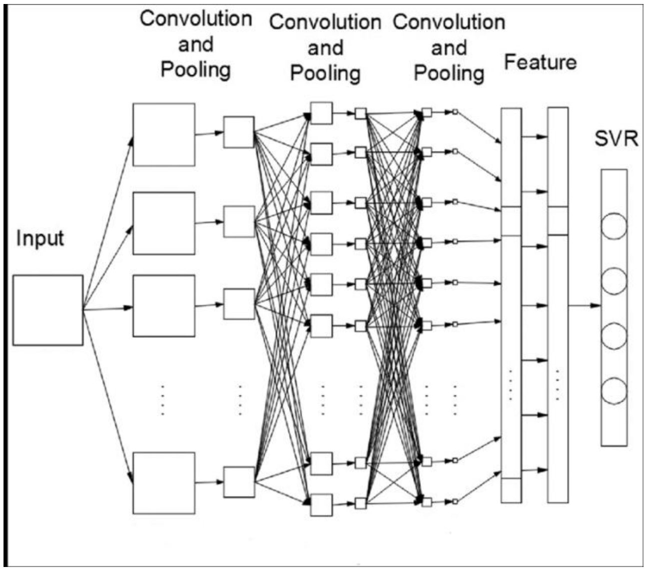

Figure 4 Internal Convolution and pooling structure of framework

The proposed architecture removes the unwanted visual features from the video frames. Then, the BS method is applied to reduce the dimension of the pixels, which will be further sent to feature extraction processing [35]. This feature extraction is computed by using CNN and later it is passed on to many hidden layers, which are shown in figure 4 along with the trained samples. Finally, linear SVM model classifies the results very accurately, when compared to the other existing methods. These normal and abnormal conditions can be predicted by using the final classifier along with an activation function. The proposed method is a combination of TBS, CNN, and SVM integrated with a softmax activation function.

5. RESULTS & DISCUSSION

Here, this research work utilizes a multi-sensor dataset named PETS, which is available in the arena dataset. In figure 5, the analysis has been performed by using deep learning method in order to predict whether the person is falling or standing. The BS method alone provides a wrong prediction output, which is later labelled. The task is predicted from all the frames of the abnormal activities detected in the sequences. The cross-validation is done by splitting the training and testing the datasets with 80% and 20% respectively [36]. Further, the proposed hybrid of TBS, CNN, and SVM is explained and examined here. Besides, testing is captured in

ISSN: 2582-2640 (online)
Submitted:6.03.2021
Revised: 30.03.2021
Accepted: 21.04.2021
Published: 10.05.2021

each folder wise. The condition can be categorized by various methods to measure the performance of the framework. Every caption on each data can be tested with various processes and it is noted. Figure 6 shows the experimental setup of the layer for training and classification procedure.

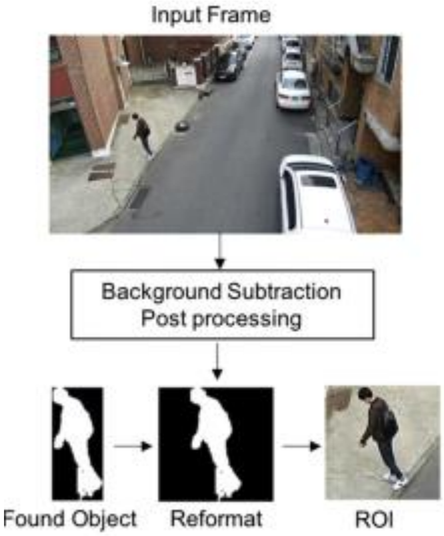

Figure 5 Experimental Results of Feature Extraction by Proposed Hybrid Model

The obtained results are tabulated accuracy-wise and ample iteration conducted for finding the efficiency of the proposed hybrid framework. This research work has used a pre-trained database for the main process to improve the accuracy of identifying the normal or abnormal condition. The classification can be done by using SVM with more accurate results, when compared to the existing procedure. Table 1 shows the performance measures of various methods. Dimension reduction is taking place in the proposed algorithm, which is notified in the table. The proposed framework is performing dimension reduction due to the combination of background subtraction with selective features from the frame of video. Therefore, the dimension can be reduced in this proposed framework.

ISSN: 2582-2640 (online)
Submitted:6.03.2021
Revised: 30.03.2021
Accepted: 21.04.2021
Published: 10.05.2021

DOI: https://doi.org/10.36548/jscp.2021.2.001

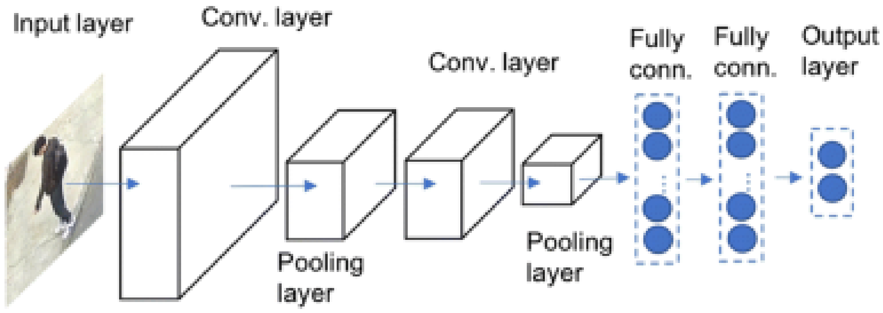

Figure 6 Experimental Setup of Classification Procedure

Table 1 Performance Measures with Existing Systems

S. No.	Methods	Classification Procedure	Real time Conditions	Results Category		Accuracy	Efficiency	DR
				20 iteration	Above 50 iteration			
1	BS	-	Manual observation needed	Normal	Normal	62%	70%	No
2	CNN	Done	Not Identified	Normal	Normal	69%	80%	No
3	SVM	Done	Partially Identified	Normal	Abnormal	72%	82%	No
4	Hybrid Proposed	Done	True Identified	Abnormal	Abnormal	90%	96%	Yes

The hybrid proposed method has proved to be very robust at many iteration conditions even to decrease the number of training datasets for classification. It is incorporated with the softmax activation function. The proposed method superiority performance is shown as a graph in figure 7.

ISSN: 2582-2640 (online)
Submitted:6.03.2021
Revised: 30.03.2021
Accepted: 21.04.2021
Published: 10.05.2021

DOI: https://doi.org/10.36548/jscp.2021.2.001

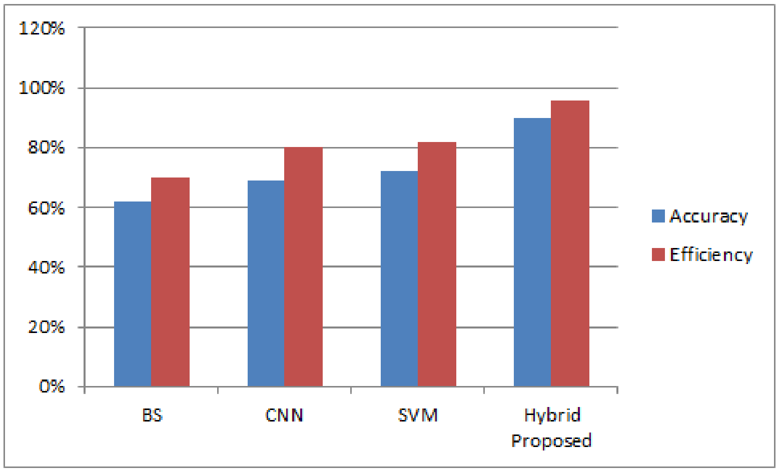

Figure 7 Performance Comparison of Proposed Framework with Existing Model

6. CONCLUSION

Thus, the proposed research article has examined several methods to classify the conditions of the video frame. The pre-processed CNN+SVM architecture provides a good accuracy with higher efficiency and less loss than other combination and single classifiers. This pre-processing is deployed for performing dimension reduction. CNN architecture remains as a very effective model to train a large number of samples with more accuracy. Besides, SVM is a very accurate classification under normal or abnormal conditions. The classifying features are mostly containing the spatial reliance in the video frames with a more temporal element. The softmax is used as an activation function in the structure of the SVM model. The proposed research work has examined this fused architecture for limited condition activities. In the future, this research work will extend to collective abnormal activities for our classification. Besides, many motion features are added for incorporating the strong pictorial features into the frame to increase the accuracy level in alliteration. This is one of the drawbacks of the current proposed framework.

REFERENCES

[1] A. A. Chaaraoui, P. Climent-Perez, and F. Fl ´ orez-Revuelta. ´ Silhouette-based human action recognition using sequences of key poses. Pattern Recognition Letters, 34(15):1799–1807, 2013.

ISSN: 2582-2640 (online)
Submitted:6.03.2021
Revised: 30.03.2021
Accepted: 21.04.2021
Published: 10.05.2021

[2] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas. Abnormal detection using interaction energy potentials. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 3161–3167. IEEE, 2011.

[3] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph Convolutional Label Noise Cleaner: Train a Plug-and-play Action Classifier for Anomaly Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1237-1246.

[4] T. Wang, M. Qiao, A. Zhu, Y. Niu, C. Li, and H. Snoussi, "Abnormal event detection via covariance matrix for optical flow based feature," Multimedia Tools and Applications, vol. 77, pp. 17375-17395, 2018.

[5] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation," IEEE Transactions on Image Processing, vol. 24, pp. 5288-5301, 2015.

[6] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," Computer Vision and Image Understanding, vol. 156, pp. 117- 127, 2017.

[7] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "An efficient subsequence search for video anomaly detection and localization," Multimedia Tools and Applications, vol. 75, pp. 15101-15122, 2016.

[8] T. Zhang, W. Jia, B. Yang, J. Yang, X. He, and Z. Zheng, "MoWLD: a robust motion image descriptor for violence detection," Multimedia Tools and Applications, vol. 76, pp. 1419-1438, 2017.

[9] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," IEEE transactions on pattern analysis and machine intelligence, vol. 36, pp. 18-32, 2013.

[10] I. Al Ridhawi, S. Otoum, M. Aloqaily, Y. Jararweh, and T. Baker, "Providing secure and reliable communication for next generation networks in smart cities," Sustainable Cities and Society, vol. 56, p. 102080, 2020.

[11] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in Proceedings of the IEEE international conference on computer vision, 2013, pp. 2720-2727.

[12] H.Wang, A. Kl¨aser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[13] P. Doll´ar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72. IEEE, 2005.

[14] D. Weinland, R. Ronfard, and E. Boyer. Motion history volumes for free viewpoint action recognition. In *Workshop on modeling People and Human Interaction (PHI'05)*, 2005.

[15] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 935-942.

[16] J. Kim and K. Grauman, "Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2921-2928.

[17] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," IEEE transactions on pattern analysis and machine intelligence, vol. 36, pp. 18-32, 2013.

[18] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in CVPR 2011, 2011, pp. 3449-3456.

[19] M. A. Alsmirat, I. Obaidat, Y. Jararweh, and M. Al-Saleh, "A security framework for cloud-based video surveillance system," Multimedia Tools and Applications, vol. 76, pp. 22787-22802, 2017.

[20] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in Proceedings of the IEEE international conference on computer vision, 2013, pp. 2720-2727.

[21] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny, "Large-scale visual relationship understanding," in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 9185-9194.

[22] C. He, J. Shao, and J. Sun, "An anomaly-introduced learning method for abnormal event detection," Multimedia Tools and Applications, vol. 77, pp. 29573-29588, 2018.

[23] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6479-6488.

[24] Y. Zhu and S. Newsam, "Motion-Aware Feature for Improved Video Anomaly Detection," arXiv preprint arXiv:1907.10211, 2019.

[25] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 341- 349.

[26] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes," IEEE Transactions on Image Processing, vol. 26, pp. 1992-2004, 2017.

[27] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in International Symposium on Neural Networks, 2017, pp. 189-196.

[28] K. Muhammad, T. Hussain, M. Tanveer, G. Sannino, and V. H. C. de Albuquerque, "Cost-Effective Video Summarization using Deep CNN with Hierarchical Weighted Fusion for IoT Surveillance Networks," IEEE Internet of Things Journal, vol. 7, pp. 4455-4463, May 2020.

[29] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 733-742.

[30] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional lstm for anomaly detection," in 2017 IEEE International Conference on Multimedia and Expo (ICME), 2017, pp. 439-444.

[31] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6536- 6545.

[32] F. Landi, C. G. Snoek, and R. Cucchiara, "Anomaly Locality in Video Surveillance," arXiv preprint arXiv:1901.10364, 2019.

[33] Y. Yu, T. Zhao, M. Wang, K. Wang, and L. He, "Uni-OPU: An FPGA-Based Uniform Accelerator for Convolutional and Transposed Convolutional Networks," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2020.

[34] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," IEEE Access, vol. 6, pp. 1155-1166, 2017.

[35] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, "Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews," Journal of computational science, vol. 27, pp. 386-393, 2018.

[36] J. Shao, C.-C. Loy, K. Kang, and X. Wang, "Slicing convolutional neural network for crowd video understanding," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5620-5628.