JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 37, 337-346 (2021)
DOI: 10.6688/JISE.202103_37(2).0004

Abnormal Crowd Behaviour Detection Using Parallel Deep Neural Networks

MUHAMMED ANEES V AND SANTHOSH KUMAR G
Artificial Intelligence and Computer Vision Lab
Department of Computer Science
Cochin University of Science and Technology
Cochin, 682022 India
E-mail: {anees; san}@cusat.ac.in

The security of people during public events is one of the significant concerns of authorities. The authorities have to monitor the entire crowd continuously, and they must be capable of preventing all the abnormal activities in the crowd. They are responsible for avoiding these kinds of situations. To prevent such abnormal behaviours, first, they need to detect the abnormal crowd behaviour from the high-density crowd under observation. Detecting abnormal crowd behaviour from the crowd video has been one of the most critical research areas in the intelligent video surveillance system field over the past decade. Numerous strategies for crowd abnormality detection with the assistance of computer vision algorithms and machine learning methods have been proposed in recent years. Many of those traditional approaches are using hand-crafted features like optical flow, HoG, SIFT, and SURF. Even though most of these methods were able to produce a considerably good performance, these methods will take a lot of computational time to extract features, and that enhances the whole computational time. Especially the high-level features like SIFT and SURF are computationally complex, and this will affect the real-time performance of the system. In this paper, we propose a novel deep learning strategy for abnormal crowd behaviour detection. Rather than utilising hand-crafted features, deep neural networks naturally learn feature representations of the crowd video and which will help the system to detect abnormal behaviours. The learned feature representations help the system to differentiate normal and abnormal crowd behaviours. This method uses convolution neural networks, which have been utilised as an integral tool for feature learning in computer vision algorithms to extract the features from the videos. Instead of using traditional single-stream convolution neural networks, we have used pre-trained two-stream convolution neural networks to detect the crowd abnormality, which can consider both spatial and temporal information in the video. Our method is tested with available standard datasets and compared with state-of-the-art methods.

Keywords: crowd flow, intelligent surveillance system, optical flow, crowd abnormal detection, crowd model, deep learning, convolution neural networks, two stream networks

1. INTRODUCTION

The utilisation of surveillance systems in public spaces has increased drastically in the last few years due to the security needs of the citizen. The utilisation of surveillance systems in public spaces has increased drastically in the last few years due to the security

Received October 30, 2019; revised February 4, 2020; accepted April 1, 2020.
Communicated by Jimson Mathew.

needs of the citizen and the increased unwanted events happening around the world [1]. Detecting and avoiding such activities is one of the primary concerns of the corresponding authorities. The intelligent surveillance system is on the most significant advancement in the field of surveillance systems, which is capable of detecting abnormalities without human intervention. Crowd analysis needs to be applied in surveillance systems to enable its intelligence. Crowd analysis is one of the critical research areas in the field of computer vision, which includes crowd monitoring, crowd tracking, action recognition, and abnormality detection.

Detection of abnormal activities from the crowded scene is one of the fundamental challenges that have to be faced while developing an intelligent video surveillance system. In recent years, many computer vision techniques are employed for detecting abnormalities from the surveillance videos. Abnormality detection from the surveillance video is challenging and complicated because it is difficult to define the abnormality. An ordinary event in a scene may be abnormal in some other views, and because of this nature, the anomaly is considered as a subjective behaviour [2]. The approaches based on traditional computer vision algorithms and machine learning may fail to understand this individual behaviour.

Traditional methods usually model standard behaviour patterns in the video scene using mathematical models and detect abnormal behaviours by considering these standard patterns. These types of approaches cannot identify all the abnormalities present in a view due to its subjective nature. In the crowd scenario, the anomalies may be formed by some rare or non-frequent movements. Identifying this unusual or non-frequent movement from the entire image or video is a massive task for a surveillance system. Most articles divided the scene into several patches and tried to detect the abnormalities in each patch to create the model. This model act as the reference model for identifying normal behaviours. If any spot does not follow this predefined reference, then it has to be considered as an outlier, and this outlier may be an unusual event. Many methods in the literature are already pursuing this strategy, which uses well-known methods like a histogram of gradients (HoG) and optical flows with some stability analysis methods [6].

In this paper, we have investigated the application of a convolutional neural network (CNN) for the detection of abnormalities from the surveillance video. Typically one stream is used in every system that uses neural networks to learn its features. Instead, here we have used two-stream convolutional neural networks that run in parallel to extract both spatial and temporal information from the crowd videos. A pre-trained VGG network was used in both streams to train the given data. The entire system is trained and tested with two public datasets. The first one is the fight dataset by VISILAB, and the second one is the web dataset released by UCF.

This paper is organised as follows. Section 2 describes the related works, and section 3 discusses the architecture of the neural network. Section 4 describes the experimental setup and result. Finally, the conclusion is given in Section 5.

2. RELATED WORK

Computer science researchers have addressed the challenges in developing an intelligent surveillance system for the last few years with the evolution of high-performance

computing devices. Most of the traditional methods in the literature use both low-level and high-level hand-crafted features like optical flow and histogram of the gradient for the abnormality detection in crowd scenes [3, 4]. Some initial works in this field use these hand-crafted features to develop some standard models to detect the crowd abnormality. Mehran *et al*. [5] proposed an abnormality detection method which uses a standard model called the social force model, and it is considered as one of the pioneering work in this field. Ali and Shah [6] proposed another model which uses Finite-Time Lyapunov Exponent (FTLE) filed for the detection of abnormalities. Krausz *et al*. [7] used the histogram of optical flow to represent the crowd behaviour patterns, and the anomaly events are detected from these motion patterns using a heuristic approach. Ragavendra *et al*. [8] proposed an abnormal event detection method using interaction forces with the help of the particle swarm optimisation algorithm. Xiong *et al*. [9] proposed an energy-based model for crowd abnormality detection using potential energy and kinetic energy.

Abnormal event detection from an unstructured crowd is more complicated than detecting abnormalities from the structured group. Wu *et al*. [10] first addresses the issue of abnormality detection from the instructed crowd using a particle advection system. The abnormal event detection system needs to consider the spatiotemporal features for the adequate representation of unusual events. Mausavi *et al*. [11] proposed a histogram-based abnormal event detection method that considers spatiotemporal movements of the crowd. The algorithms we have listed till now discusses only the hand-crafted features for modelling the crowd flow. But recently proposed methods are the combination of both spatiotemporal features and textual information. Li *et al*. [12] make use of both the textual information and spatiotemporal features for the detection of abnormal events. Kalsta *et al*. [13] proposed an abnormal behaviour method with the help of a Histogram of Swarms (Hos) as textual information and Histogram of Gradient as the feature (HoG). These methods mentioned above work on the features extracted from the videos. However, these methods have some limitations and drawbacks. Most of these features are designed to use in general-purpose images, and we are trying to incorporate those features into our intelligent surveillance system. Degradation of performance resulted while applying those feature extraction mechanisms in videos. The second biggest challenge is the selection of an appropriate feature for the scenario. If we are selecting a wrong feature for developing the system for a particular scene, then the entire system may fail.

Recently Convolution neural network is proved to be effective in addressing many challenging problems in various fields, such as problems in image processing based on classification like image classification [14], object detection [15], and activity recognition [16]. The approaches based on CNN can perform significantly better than traditional methods. Traditional hand-crafted feature-based methods are computationally more complex than deep neural networks. Loris Nanni *et al*. [17] has conducted a comparative study between Hand-crafted features and convolution neural networks and claims that CNN is computationally inexpensive compare to the hand-rafted features. Due to these advantages, we are using a convolution neural network for crowd abnormal detection. Feng *et al*. [18] introduced a deep learning framework based on deep Gaussian Mixture Model (GMM), the detection abnormality from the high-density crowd. Sabokrou *et al*. [19] proposed a crowd abnormality detection method that uses convolution neural networks completely. Wei [20] improved this method using a modified deep neural network called a two-stream, fully convolution neural network. Lazaridis *et al*. [21] proposed a

heat map based on two-stream neural networks to detect the crowd anomaly. These are the latest works available in the literature.

In this proposed method, we have implemented the crowd abnormality detection using a Pre-trained two-stream VGG16 network. VGG network is used in this scenario due to its performance and robustness [22]. VGG has a total of 138 million parameters to achieve its performance. The convolution kernel size of VGG is 3×3, and the max pool size is 2×2, which helps the network to learn more precise and fine features. The obtained results are also compared with the milestone works available in the literature.

3. ARCHITECTURE

The architecture of the newly proposed crowd abnormality detection system is given in Fig. 1. Emerging deep learning techniques are used in our work to detect abnormal crowd videos. Instead of using the single-stream convolution neural network, 2-stream convolution neural networks are adopted to identify the abnormalities. A single stream of layers is usually used in deep learning techniques for implementing convolution neural networks. But in this work, two parallel streams of systems are used to learn more features than usual single stream networks. The raw frames taken from the input video is used as the first stream, and optical flow frames are used as the second stream. Stream one is used to learn the appearance information or spatial information from the regular frames extracted from the input video, and stream two is used to acquire the motion information or temporal information from the corresponding optical flow frame.

The crowd videos are given to the proposed abnormal detection system as input. Videos can be considered as the continuous flow of image frames so that we can extract a fixed number of frames from the input video. We have used raw frames obtained from the input video for training the first stream for the learning purpose of spatial information. Before feeding into the training module, some preprocessing steps have to be applied. In the preprocessing steps, the extracted frames are resized into 299×299 pixels. These resized frames are fed into a deep learning framework for training. In this proposed method, we are using a pre-trained VGG -16 network [22] for training purposes.

VGG is a convolution neural network proposed by the Visual Geometry Group of the University of Oxford. VGG has two versions, VGG 16 and VGG 19. VGG 16 is a pre-trained model with 16 weight layers, whereas VGG 19 consists of 19 weight layers.The size of input layer in VGG is $224 \times 224 \times 3$. The input layer to the last max-pooling layer is the feature extraction part, and the rest of the network is considered as a classification part of the model. Softmax layers are usually used for the final classification with the respective number of classes. In our problem, the output layer is a softmax layer with two classes.

The first Convolution stream gives us the first convolution model, and this model is created based on the real images extracted from the supplied video input. This model can represent only the spatial information in the video, but it cannot describe the temporal information in the scene. We need both spatial and temporal information of the data to analyze the exact crowd behaviour from the input crowd video. To represent the temporal information, the second convolution stream, which takes optical flow frames as input.

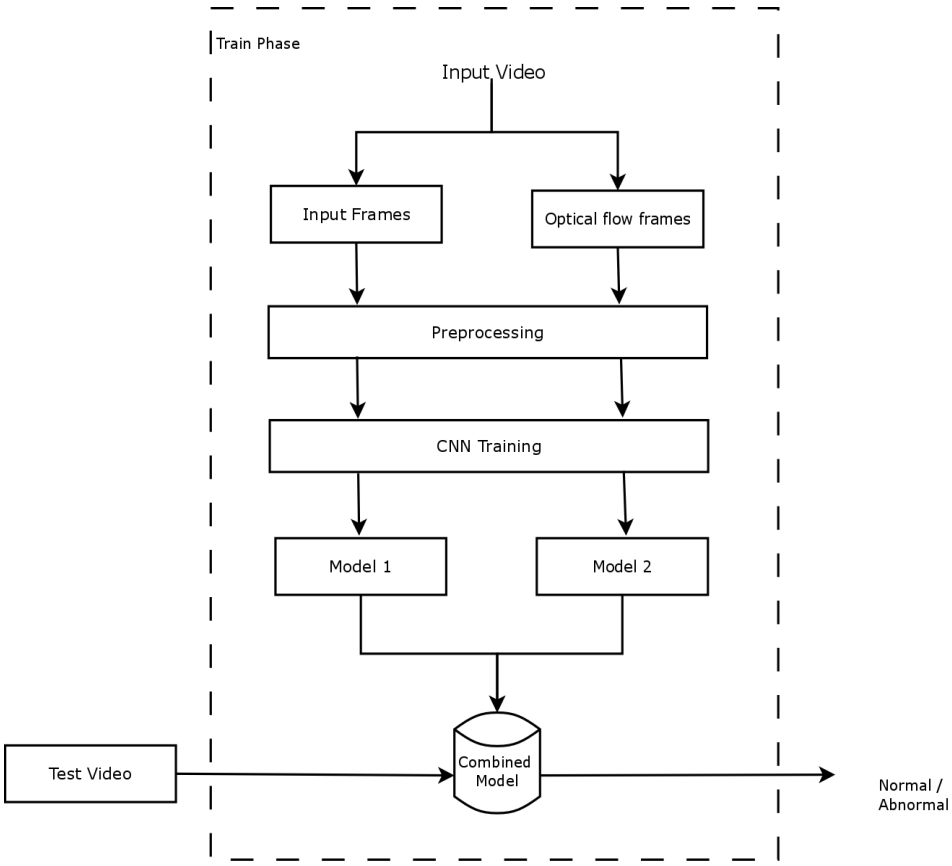

Fig. 1. Proposed system architecture.

Optical flow is a 2D motion vector that shows the motion pattern resulted from the movements between two consecutive frames. We are using Gunner Farneback optical flow [23] algorithm to extract optical flow frames from the input crowd video. Gunner Farneback is used to calculate the dense optical flow present in the video. Sparse techniques require only some pixels to process from the whole image, whereas dense techniques process entire pixels. Dense methods are slower than sparse but are more accurate. The extracted optical flow frames are fed to the same VGG-16 ntwrk to create the second stream, which is explained above. The second stream outputs the second convolution model, which can represent both the spatial and temporal contents of the video.

Now we have two convolution models obtained from two independent convolution streams running in parallel. The spatial model and the temporal model are merged to create a new convolution model to incorporate both the temporal and spatial features. This new model contains the properties of both the spatial model and the temporal model, and this combined model can represent both spatial and temporal information of the data. So we obtained the combined model from the raw frames and optical flow frames, and the coupled model can detect the abnormal events present in the video.

4. EXPERIMENTAL RESULTS AND ANALYSIS

This section subsumes the experimental setup and analysis of the proposed model. The fight dataset [24] released by VISILAB and web dataset by UCF [25] are used here. These datasets are released specifically for evaluating and measuring the crowd abnormality detection systems. The fight dataset considers both normal and abnormal events occurring in similar, but dynamic conditions. The UCF web dataset also contains both normal and abnormal videos from different scenarios.

This fight dataset by VISILAB contains two sets of videos. The first video set is collected from the National Hockey League (NHL) of Spain, and the second set is collected from the various action sequence of different movies. This dataset contains a total of 246 videos from both normal and abnormal categories. One hundred and twenty-three videos fall under the normal category, and the remaining 123 videos fall under the abnormal category. Each video in this dataset contains 50 frames of size 720×576. Ninety-eight videos from each class are taken for training, and the remaining 25 videos are taken for testing both normal and abnormal class.

The UCF web dataset contains videos from different sources. The dataset includes 17 abnormal videos and 27 unusual videos. The videos from various sources with different resolution and different aspect ratios which differ in length are included. Thirty among these are included in training, and 14 are used for testing. So a total of 228 videos are fed for training, and 64 videos are taken for testing. The entire system runs for 200 epochs in the training phase.

We have tested our dataset with both single-stream network and double stream networks. We have passed the training data into the VGG network to create the first convolution model. Then the optical flow frames are extracted from the training videos and moved to the same VGG network to create the second model. Then these two models are combined to get a more refined and accurate model for detecting the abnormalities from the crowd video. Then the model is tested with our test data set that contains 64 test videos, fifty from the fight dataset, and 14 from the web dataset. 20 out of 25 abnormal videos in fight dataset are classified as unusual, and 24 videos out of 25 normal videos are categorised as usual using the two-stream model. The F1-score matrix created from these results is given in Table 1.

Table 1. F1-score matrix for fight dataset.

	Actual Normal	Actual Abnormal
Predicted Normal	TP = 24	FP = 5
Predicted Abnormal	FN = 1	TN = 20

Fourteen videos from web dataset are used for testing. Eight videos are taken from normal class, and six videos are from abnormal class. Seven normal videos are classified as normal, and one is wrongly classified as abnormal. All the six videos out of 6 abnormal videos are correctly classified. The F1-score matrix for the above results is given in Table 2.

The combined F1-score matrix is calculated and given in Table 2. Thirty one normal videos are correctly classified as normal videos, and twenty six abnormal videos are cor-

Table 2. F1-score matrix for web dataset.

	Actual Normal	Actual Abnormal
Predicted Normal	TP = 7	FP = 0
Predicted Abnormal	FN = 1	TN = 6

rectly classified as abnormal behaviour. Two normal videos and five abnormal videos are wrongly classified.

Table 3. F1-score matrix for the entire data.

	Actual Normal	Actual Abnormal
Predicted Normal	TP = 31	FP = 5
Predicted Abnormal	FN = 2	TN = 26

We can calculate precision, recall, F1-score, and Accuracy from the F1-score matrix. We got precision as 86.11% , Recall as 93.94%, F1-score as 89.86% and Accuracy as 89.06%. We compared the two-stream CNN results with both single-stream networks. We have tested our dataset with both spatial and temporal models. The spatial model gives us the precision as 81.48% , Recall as 88.00%, F1-score as 85.62% and Accuracy as 84.00%. The temporal model gives us the the precision as 85.19% , Recall as 92.00%, F1-score as 88.46% and Accuracy as 88.00%. The results are tabulated in Table 4.

Table 4. Comparison of single stream and two stream networks.

	Precision	Recall	F1-Score	Accuracy
Spatial Stream CNN	81.48	88.00	85.62	84.00
Temporal Stream CNN	85.19	92.00	88.46	88.00
Two Stream CNN	86.11	93.94	89.86	89.06

From the above table, it is clear that two-stream convolution networks can outperform both single-stream systems. The entire system is implemented using OpenCV and Keras libraries in Python. The experiment is executed in a system with NVIDIA K80 GPU.

The proposed model is compared with existing state-of-art models in the literature. Most of the works in the literature are based on single-stream convolution networks that take raw input frames. All those methods are tested with our dataset, and the comparison results are as shown in Table 5. From this figure, we can understand that our method can outperform all the crowd anomaly detection methods in the literature.

The traditional convolution model is not capable of distinguishing between spatial and temporal information. These models lack motion-sensitive information while learning from the video data. In our model, motion information is given explicitly to the system in the form of optical flow vectors. All the normal features are learned from the original image, and motion features are learned from optical flow images. So the combined model is capable of addressing both elements, and it helps the system to perform better than other reported works.

Table 5. Comparison study.

state of art methods	Accuracy
Sabokrou *et al*.	69 %
Dan Xu *et al*.	75 %
Spatial	84 %
Lazaridis	86%
Temporal	88 %
Wei's method	88 %
Proposed method	89 %

5. CONCLUSION

In this paper, we proposed a deep learning framework for abnormality detection from crowd videos. Instead of using a single-stream neural network, we used two-stream neural networks for abnormality detection. The proposed two-stream CNN uses a pre-trained VGG-16 model for model creation. The model is tested with a fight dataset released by VISILAB and the web dataset released by UCF. The experimental results show an accuracy of 89% for two streams CNN for the combined dataset. Our proposed two-stream model for crowd abnormality detection outperforms all the existing crowd models in the literature on our combined dataset, and this model can be extended to real-time applications. The performance of the system depends on the quality of the optical flow frames, and Some preprocessing might be helpful for the system to learn more accurate features and it is an exciting direction for future research.

REFERENCES

1. Y. Zhang, H. Lu, L. Zhang, and X. Ruan, "Combining motion and appearance cues for anomaly detection," *Pattern Recognition*, Vol. 51, 2016, pp. 443-452.
2. M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette, "Real-time anomaly detection and localization in crowded scenes," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Vol. 2015, 2015, pp. 56-62.
3. Y. Zhang, L. Qin, H. Yao, P. Xu, and Q. Huang, "Beyond particle flow: Bag of trajectory graphs for dense crowd event recognition," in *Proceedings of IEEE International Conference on Image Processing*, 2013, pp. 3572-3576.
4. Y. Cong, J. Yuan, and J. Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognition*, Vol. 46, 2013, pp. 1851-1864.
5. R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 935-942.
6. S. Ali and M. Shah, "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1-6.

7. B. Krausz and C. Bauckhage, "Analyzing pedestrian behavior in crowds for automatic detection of congestions," in *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2011, pp. 144-149.
8. R. Raghavendra, A. D. Bue, M. Cristani, and V. Murino, "Optimizing interaction force for global anomaly detection in crowded scenes," in *Proceedings of IEEE International Conference on Computer Vision Workshops*, 2011, pp. 136-143.
9. G. Xiong, X. Wu, Y. Chen, and Y. Ou, "Abnormal crowd behavior detection based on the energy model," in *Proceedings of IEEE International Conference on Information and Automation*, 2011, pp. 495-500.
10. S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2054-2060.
11. H. Mousavi, S. Mohammadi, A. Perina, R. Chellali, and V. Murino, "Analyzing tracklets for the detection of abnormal crowd behavior," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2015, pp. 148-155.
12. H. Guo, X. Wu, N. Li, R. Fu, G. Liang, and W. Feng, "Anomaly detection and localization in crowded scenes using short-term trajectories," in *Proceedings of IEEE International Conference on Robotics and Biomimetics*, 2013, pp. 245-249.
13. V. Kaltsa, A. Briassouli, I. Kompatsiaris, L. J. Hadjileontiadis, and M. G. Strintzis, "Swarm intelligence for detecting interesting events in crowded environments," *IEEE Transactions on Image Processing*, Vol. 24, 2015, pp. 2153-2166.
14. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of ACM*, Vol. 60, 2017, pp. 84-90.
15. S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, Vol. abs/1506.01497, 2015.
16. K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *CoRR*, Vol. abs/1406.2199, 2014.
17. L. Nanni, S. Ghidoni, and S. Brahnam, "Handcrafted vs. non-handcrafted features for computer vision classification," *Pattern Recognition*, Vol. 71, 2017, pp. 158-172.
18. Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, Vol. 219, 2017, pp. 548-556.
19. M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Computer Vision and Image Understanding*, Vol. 172, 2018, pp. 88-97.
20. H. Wei, Y. Xiao, R. Li, and X. Liu, "Crowd abnormal detection using two-stream Fully Convolutional Neural Networks norm single frame norm optical flow norm," in *Proceedings of the 10th International Conference on Measuring Technology and Mechatronics Automation*, 2018, pp. 332-336.
21. L. Lazaridis, A. Dimou, and P. Daras, "Abnormal behavior detection in crowded scenes using density heatmaps and optical flow," in *Proceedings of the 26th European Signal Processing Conference*, 2018, pp. 2060-2064.
22. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, Vol. abs/1409.1556, 2014.

23. G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proceedings of the 13th Scandinavian Conference on Image Analysis*, 2003, pp. 363-370.
24. E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Proceedings of the 14th International Conference on Computer Analysis of Images and Patterns*, Vol. part 2, 2011, pp. 332-339.
25. R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model." in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 935-942.

Muhammed Anees V completed his post-graduation in Computer Science from Cochin University of Science and Technology. He is currently working as a Research Scholar in the Department of Computer Science, Cochin University. His research interests include computer vision and image processing.

Santhosh Kumar G earned his Ph.D. degree in Computer Science from Cochin University of Science and Technology (CUSAT). He is currently working as a Professor in the Department of Computer Science, CUSAT. His research interests include wireless sensor networks, cyber-physical systems, and computer vision.