ANOMALY AND ACTIVITY RECOGNITION USING MACHINE LEARNING APPROACH FOR VIDEO BASED SURVEILLANCE

Aiswarya Mohan[1], Meghavi Choksi[2] and Mukesh A Zaveri[3]
Computer Engineering Department,
Sardar Vallabhbhai National Institute of Technology, Surat, India-395007
[1]aiswaryamohanpulikkal@gmail.com, [2]meghavichoksi@gmail.com and [3]mazaveri@coed.svnit.ac.in

***Abstract*—In the current era, the majority of public places such as supermarket, public garden, mall, university campus, etc. are under video surveillance. There is a need to provide essential security and monitor unusual anomaly activities at such places. The major drawback in the traditional approach, that there is a need to perform manual operation for $24 * 7$ and also there are possibilities of human errors. This paper focuses on anomaly detection and activity recognition of humans in the videos. The anomaly detection system uses principal component analysis network (PCANet) and Convolutional Neural Network (CNN) to solve the problems of manual operation such as the false alarms, missing of anomalous events and locating the position of an anomaly in the video. The frames wise abnormal event is detected using principal component analysis and Support Vector Machines (SVM) classifier. The location of the abnormality in a frame is detected using Convolutional Neural Network.**

***Index Terms*—Video Surveillance, Activity Recognition, Machine Learning, Anomaly Detection.**

I. Introduction

A lot of research has been proposed in the field of video surveillance in terms of object tracking [1]–[3] and human behavior [4]–[7]. There are numerous applications of video surveillance in various fields such as crowd surveillance, industrial monitoring, forest fire controlling, traffic surveillance, aerial monitoring, security surveillance, post disaster management etc. Issues with conventional video surveillance approaches are: (i) lack of real time video processing, (ii) time consuming, (iii) human error that leads to a false alarm, (iv) maintenance and storage constrain, (v) inefficient when there is a large crowd. Manual operation for action recognition is time consuming, tiresome, and inefficient especially over a place where the crowd is dense. There is a need of an automated system which optimizes operational issues and sends alerts to human operators without a time delay. This system should be capable to detect operational errors and generates a notification. The system should be capable to track an abnormal event in each frame and generate a notification of such an event. An object tracking and anomaly detection can be improved by applying machine learning techniques on it.

Machine learning [8], [9] provides the ability to learn from a trained dataset. The video surveillance application such as abnormal event detection or activity recognition provides better results by using machine learning. For instance, one of the most standard techniques such as CNN [10] has been designed to handle even the three-dimensional data of the video input. For example, CNN can be used to extract the features of an image, classify images etc. CNN can extract the features and patterns in a video faster than traditional image processing techniques. The standard Support Vector Machines technique for abnormal event performs poorly when used independently whereas, when used with classification after converting features using deep learning algorithm it performs far better [11].

There are numerous events such as music concerts, protests, festivals, sports events, etc where a large crowd is gathered. During such events, there is a need of real time and an intelligent system that can help to detect abnormal activities. The system designed for crowd monitoring at the public spaces like train stations, stadiums, airport terminals, theaters should consider crowd density, peak time of traffic in the areas etc. The proper crowd management at public places can help to avoid crowd mismanagement and ensure public safety. Various deep learning algorithms can be employed to analyze the behaviour of the crowd from features like optical flow and the changes of optical flow to predict the chance of occurrence of disaster in the crowd in real time [7], [12]. The outlined of paper is as follows: related work is described in Section 2, the proposed approach is mentioned in section 3. In section 4, simulation results are explained. Section 5 describes a conclusion, it is followed by an acknowledgment and references.

II. Related Work

Crowd surveillance has been the most explored field in recent decades, lots of discussions and research studies have been done [4], [7], [13], [14]. The recent literature targets in the area of crowd surveillance such as anomaly event detection, object tracking, person tagging etc, where machine learning has been applied. The conventional model for crowd behavior is an agent based model [15], flow based model [11] and particle based model. To represent pedestrians like a continuous density field in space and also have proposed partial differential equations to describe the dynamics of a crowd is developed by Hughes et al. [13]. A real time crowd modelling

based on continuum dynamics has been presented by Treuille et al. [4] describe individuals motion using dynamic potentials and velocity fields. In the agent-based model, individuals or group in the crowd are represented as an agent have decision making capability and have awareness of their surrounding environment and works on a set of predefined rules. A generalized form of the social force model has been used in a many of simulation problems in crowd behaviour analysis and computer graphics. The social force model proposed by Helbing et al. [5] uses the agent based approach, it considers the crowd motion and environmental constrain.

Human activities in a crowd motion are classified based on features into: (i) optical flow based features, (ii) trajectory or tracklet and (iii) local spatio-temporal features. The flow based features calculate directly from pixel values. In a general scenario crowd at public events where the crowd is more irregular, so in such cases, the trajectories based approach would not work well. The flow based method avoid tracking at the macroscopic level. Tracklets and trajectories have been used when the crowd is regular, repetitive and crowd density is low [16]. In an outdoor scenario, the flow based features are more suitable with wide field view and low resolution of the targets and when the aim is to analyze a holistic trend of the crowd. Local spatio-temporal features are calculated from 2D or 3D patches from each frame in a video. Local spatio-temporal features are useful in scenarios where flow features fail, like in crowded scenes, where the motion and crowd is not uniform and motion is generated by a number of moving objects.

Developments in deep learning in the field of physical security and surveillance is path-breaking. In the field of video surveillance, several applications or problem statements stand out that can benefit from machine learning and deep learning, abnormal event detection and activity recognition being two examples. Deep learning approaches can improve the image processing technique efficiently without explicitly processing the features like edges, RGB levels etc. of an image separately [14], [17]. Though traditional machine learning approaches like SVM, regression learning, are used in an image processing and enhancement field. Tenserflow [18] provides significant success in predicting, creating, classifying an object or activities by using its various model.

The novelty of the proposed approach is the combination of anomaly detection at the frame level with that at the sub-frame level. While detecting which frames have activity different from the baseline it is equally important to detect which part of the frame has the anomalous activity happening in and to classify the activity. The motivation to do this work is the idea that Machine Learning can improve the results of video surveillance in terms of accuracy and cost. The cost can be reduced in terms of less manual operation, less storage for processing as compared to traditional image processing techniques. Automated detection of bnormal events and recognition of activities of humans in the video can be helpful.

Fig. 1. Block diagram of an anomaly activity recognition system

III. System Framework and Model

The paper focuses on human activity recognition in a crowd monitoring system, it recognizes the action of a human and detects anomaly events in real time with the minimum delay. The proposed system uses the CNN model to detect abnormality in the frame. Figure 1, represents a basic block diagram of the anomaly activities or event in a surveillance system. The main objective of crowd surveillance is to detect and report any abnormal activities or happenings in the area under monitoring. Abnormal activities or events in a video are the occurrences of that are unusual events happens in an irregular behaviour [19]. An efficient video surveillance system should be able to detect and identify an object. An object can be a person as well as group of person even the trickier ones like bikes, cars, cats, dogs and packages, etc. These detected objects are further tracked and processed in the consecutive frame. To track the objects between multiple cameras feeds from the point in time of selection. To recognise the action of the object in our case object is a human being in the video. For object recognition, machine learning technique is applied on the training data set. It detects the occurrence of an anomaly event in the scene under observation and it should generate alerts without a time lag.

Figure 2 represents the block diagram of the system designed for abnormal event detection. The solution proposed here is to find the abnormal frames first and then process these abnormal frames to locate the abnormal activity. For the first stage, the features of saliency information (SI) and multi-scale histogram of optical flow (MHOF) are used. For better performance Multi-scale histogram of optical flow [20] processes the video patch-wise, where each frame is divided into a number of image patches. This reduces computational complexity. Principal component analysis (PCA) is performed on the extracted features to obtain higher level features. An SVM model is obtained by training the extracted features in SVM. The method of feature extraction and model formation for frame-wise abnormal event detection is described in the Algorithm 1. Multi-scale histogram optical flow to preserve more precise extracted frame information as compared to the traditional histogram of optical flow (HOF) [20]. MHOF is capable to detect a change in the current frame and helps to detect abnormal events in consequent frames accurately. The equations 1 and 2 are used in the algorithm to calculate the class-label of each pixel class i,j where TH is the threshold

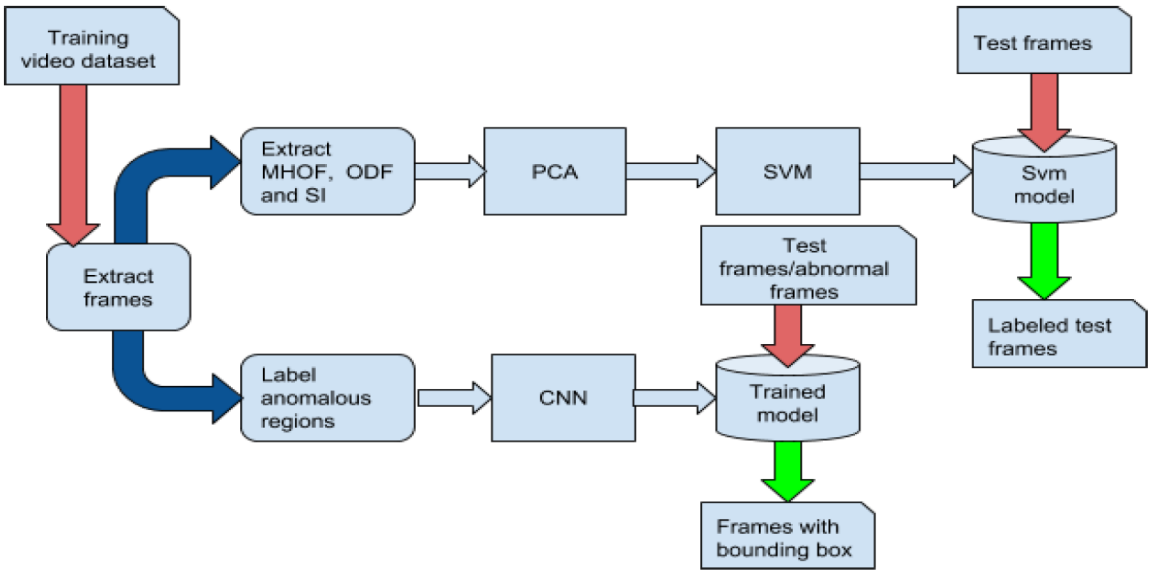

Fig. 2. Block diagram for abnormal event detection

[20].

$$C_{ij}\varepsilon \begin{cases} 0, & d^x_{i,j},\, d^y_{i,j} \leq \text{TH} \\ 1, & d^x_{i,j},\, d^y_{i,j} > \text{TH} \end{cases} \quad (1)$$

where, the optical flow of each pixel is represented by $(d^x_{i,j}, d^y_{i,j}$); the magnitude threshold is represented by TH; θ is the angle of $d^x_{i,j}$,$d^y_{i,j}$.

$$class_{i,j} = round(\theta(d^x_{i,j}, d^y_{i,j})/(\pi/4)) + 8 * C_{i,j} \quad (2)$$

After the model is formed the new set of frames are processed, for abnormal event detection, it is given as an input to the SVM model. The model classifies the images as abnormal or normal frames. Now, to locate the abnormality within the frame. Location of the abnormalities within a frame is detected by CNN. A pre-trained model i.e. represented by model_i can be retrained with the dataset of abnormal event detection to form a new model i.e. represented by model_j. The model_j is now a CNN model that can locate anomaly in a frame and gives an output of bounding box in frame to point region of an anomaly in the frames initially classified as abnormal by the Algorithm 1. All these frames are then parsed to stream as video output. In step 5 of Algorithm 2 the image along with the features of the bounding box, i.e, coordinates of the bounding box are input on to the CNN.

The CNN takes the images as 3D input volume consisting: its pixel values along the two dimensions value at each RGB layers. Inside the CNN the weights are configured according to the new input data set to form the retrained model that has structure as shown in Figure 3, we use a CNN model to train, detect and annotate the action of a person in a test video. The method of action recognition is as described in Algorithm 3. A pre-trained CNN model can be retrained for action recognition. The pre-trained model is trained with the new labeled data set, which here are the frames of videos of individual humans doing different activities like running, biking etc. The retrained model can now be used for classification of the activity of the human, in the input test video as one of the classes the system is trained. Figure 3 training dataset is input into the pre-trained model. This dataset includes the frames of training videos along with their labels identifying the action of the

Algorithm 1 Model creation

1: **Input:** Set of the frames F where, F is $\{f_1, f_2, \ldots, f_n\}$, where n is the total number of the frames and f_i is the i_{th} frame in the video.
2: **Output:** Trained SVM Model.
3: For each frame f_i
4: **repeat**
5: Divide image into $m * n$ patches.
6: Extract S the saliency feature of i. S is $\{s_i, s_2, \ldots s_k\}$, where $k = m * n$ is the patches into which image x is divided. $S_i = \sum_{i!=j} W_{i,j} * D_{i,j}$, where $D_{i,j}$ is the difference between patches i and j, $w_{i,j}$ is the corresponding weight difference.
7: Extract MHOF value H of the frame according to equation 1 and 2. H_i is the MHOF of frame i. $H_i = \{h_1, h_2, \ldots h_k\}$ where H_j is the MHOF of the patch
8: Extract ODF value HH of the frame. HH i is the HOF of the frame i. $HH_i = \{hh_1, hh_2, \ldots hh_k\}$ where h j is the ODF of patch j
9: **until**
10: Form training data set $((S_1, H_1, HH_1)\ (S_2, H_2, HH_2\ ), \ldots, (S_n, H_n, HH_n))$
11: Using principal component analysis transform training_data to data
12: Train SVM using data to obtain svm model.

Algorithm 2 Abnormal event detection- locating abnormal region

1: **Input:** F the set of frames for training and T the testing set of frames.
2: **Output:** T labeled with bounding boxes
3: For each frame f in the F manually label region of abnormality.
4: Obtain ymax, ymin, xmax and xmin the coordinates of corners of bounding box.
5: Input the images of F and the x and y values obtained from step 7 of Algorithm 1 to CNN pre-trained model retrain the model.
6: Input frames starting from first abnormal frame in T detected using Algorithm 1 to the retrained model.

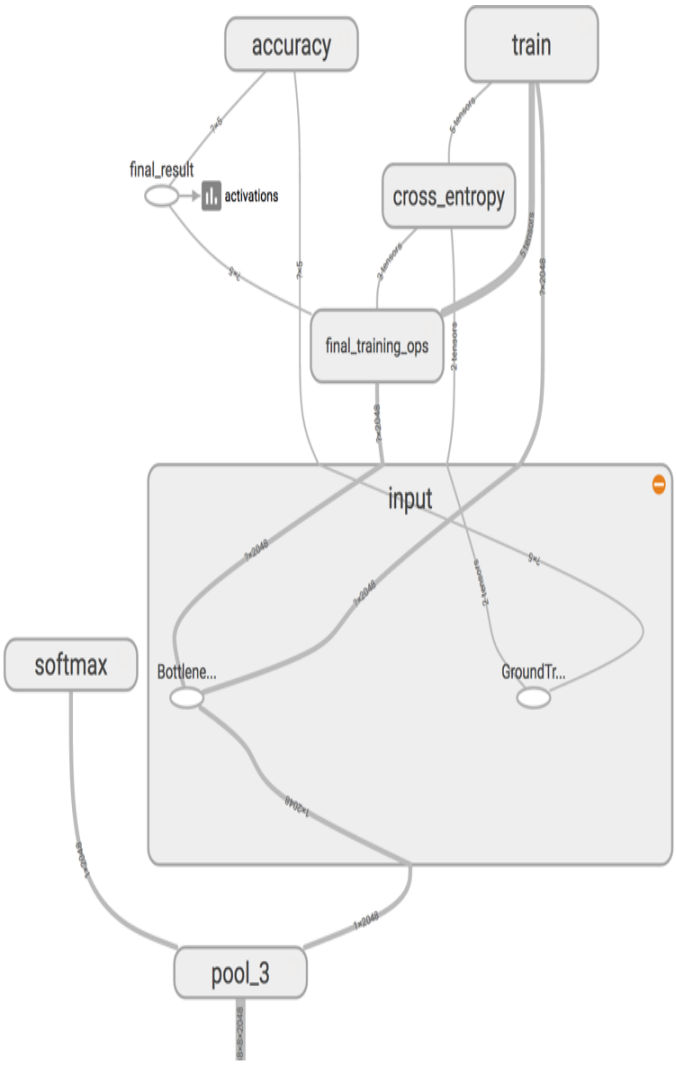

Fig. 3. Re-trained CNN model

person in a frame.

The newly generated model is further processed for anomaly detection. After running Algorithm 1 and Algorithm 2 on the data the output is a set of anomalous frames. These frames are given as an input to Algorithm 3. The Algorithm 3 recognized anomaly actions in the frame and labeled those frame.

Algorithm 3 Action Recognition

1: **Input:** Frames F $\{f_1, f_2, \ldots, f_n\}$, where f_i is set of frames with label i, F frames of video to be labeled
2: **Output:** Frames labeled by the action of the human in the video
3: Input F to pre-trained model i.e. model_i
4: The model after transfer learning is obtained model_j
5: Input F to model_j to obtain labeled output, i.e, frames labeled with classes

IV. Result and Analysis

The result and analysis of the proposed algorithms are described in this section. The tools used for the study are Tensorflow framework [18] for deep learning which is installed on a GPU system and MATLAB R2014a. The algorithm is evaluated with three datasets in terms of F1 score and accuracy. For evaluation of an anomaly detection we have used three datasets: (i) UMN Dataset [21], (ii) Avenue Dataset [22] and UCSD Dataset [23] and shown comparison in terms of F1 score and accuracy. For action recognition the data set is formed by using subsets of UCF-101 [24] action recognition dataset [22].

TABLE I
FRAMEWISE DETECTION PERFORMANCE WITH UMN AND AVENUE DATASET IN F1 SCORE

Dataset	**Proposed PCANet with MHOF, ODF and SI**	**PCANet with SI and MHOF [20]**
UMN Dataset [21]	0.99978	0.99543
Avenue Dataset [22]	0.98120	0.98737

The frame-wise performance for abnormal event detection, using UMN and Avenue dataset is as shown below in Table I. Also in Table I, an existing system [20] is compared with the proposed approach. As shown in Table I for UMN dataset, the proposed MHOF, ODF and SI features along with PCANet for feature extraction and for classification the performance in terms of F1 score is 0.99978, while for an existing system it is 0.99543. Similarly, for avenue dataset, the proposed approach provides F1 score of 0.98120. Also, the proposed approach for UMN dataset is compared with Bayesian Net with MHOF and SI, F1 score obtained in the existing Bayesian Net is 0.99950. For Avenue dataset we have compared with MHOF and SI with SVM, F1 score obtained is 0.98105. Figure 4 shows an example of anomaly detected and it is marked by a red bounding box.

Fig. 4. An anomaly detected in Avenue dataset

Abnormal event detection using CNN

Accuracy of an abnormal event detection of the proposed approach using CNN for a different dataset is shown in Figure 5. We have compared with three dataset namely Avenue [22], UCSD [23] and UMN [21]. The proposed approach is evaluated on all 3 datasets and it is observed that accuracy of avenue dataset performs better comparatively as shown in Figure 5. Avenue dataset has 16 training videos and 21 test videos. UCSD dataset has data in two subsets Ped1 and Ped2 with about 200 frames in each video. From each dataset, around 10 videos, of testing subset was tested to attain these results.

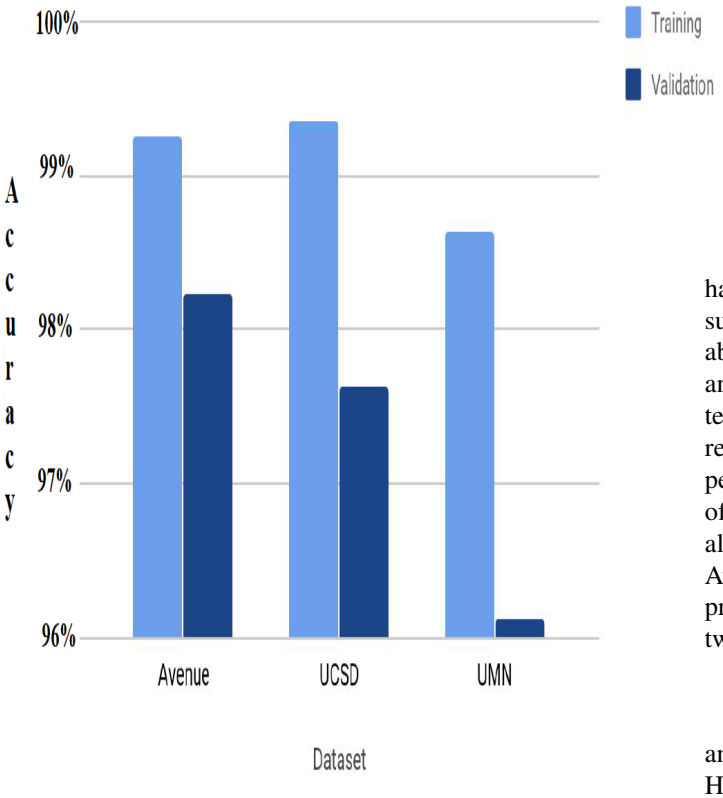

Fig. 5. Classification accuracy for different datasets

Action Recognition

For action recognition, about 100 images from Google image search results were tested in addition to the subframes detected from the anomaly detection system. The proposed action recognition algorithm as mention in Section III, detect an anomaly action is compared with CNN (framewise classification) and Multilayer Perceptrons (MLP) in terms of Top 1 accuracy and Top 5 accuracy, as shown in the Table II. It is observed that the proposed approach performs better and gives Top 5 accuracy of 94%.

TABLE II
ACTION RECOGNITION ACCURACY

Algorithm	**Top 1 Accuracy**	**Top 5 Accuracy**
CNN (frame wise classification)	69%	92%
CNN to extract features + Recurrent Neural Networks (RNN)	71%	94%
CNN + Multilayer Perceptrons (MLP)	70%	88%

V. Conclusion

The growing demand for a secure and safe environment has enhanced the research for developing smart automated surveillance systems. These systems are expected to be adaptable, dynamic, reliable as well as affordable. The proposed anomaly and activities recognition system automatically detects anomaly events and notifies with a tag. The action recognition system implemented classifies the actions of a person in the video with an accuracy of Top 1% accuracy of 71% and Top 5% accuracy of 94%. Also, the proposed algorithm is evaluated in terms of accuracy with three datasets: Avenue, UCSD and UMN, it is observed that proposed approach performs better in Avenue dataset as compare to other two.

Acknowledgment

This work is supported by the Ministry of Electronics and Information Technology (MeitY), funded by Ministry of Human Resource Development (MHRD), Government of India (Grant No. 13(4)/2016-CC&BT).

REFERENCES

[1] S. Ojha and S. Sakhare, "Image processing techniques for object tracking in video surveillance-a survey," in *Pervasive Computing (ICPC), 2015 International Conference on*. IEEE, 2015, pp. 1–6.

[2] K. A. Joshi and D. G. Thakore, "A survey on moving object detection and tracking in video surveillance system," *International Journal of Soft Computing and Engineering*, vol. 2, no. 3, pp. 44–48, 2012.

[3] G. L. Foresti, "Object recognition and tracking for remote video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 7, pp. 1045–1062, 1999.

[4] A. Treuille, S. Cooper, and Z. Popović, "Continuum crowds," *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3, pp. 1160–1168, 2006.

[5] A. Johansson, D. Helbing, H. Z. Al-Abideen, and S. Al-Bosta, "From crowd dynamics to crowd safety: a video-based analysis," *Advances in Complex Systems*, vol. 11, no. 04, pp. 497–527, 2008.

[6] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 31, no. 3, pp. 539–555, 2009.

[7] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification," in *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*. IEEE, 2017, pp. 1–7.

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002, pp. 79–86.

[9] M. Narwaria and W. Lin, "Svd-based quality metric for image and video using machine learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 347–364, 2012.

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," pp. 1–9, 2017.

[11] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: A survey," *IEEE transactions on circuits and systems for video technology*, vol. 25, no. 3, pp. 367–386, 2015.

[12] M. S. Zitouni, H. Bhaskar, J. Dias, and M. E. Al-Mualla, "Advances and trends in visual crowd analysis: a systematic survey and evaluation of crowd modelling techniques," *Neurocomputing*, vol. 186, pp. 139–159, 2016.

[13] R. L. Hughes, "A continuum theory for the flow of pedestrians," *Transportation Research Part B: Methodological*, vol. 36, no. 6, pp. 507–535, 2002.

[14] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2871–2878.

[15] V. Kountouriotis, S. C. Thomopoulos, and Y. Papelis, "An agent-based crowd behaviour model for real time crowd behaviour simulation," *Pattern Recognition Letters*, vol. 44, pp. 30–38, 2014.

[16] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 5, pp. 1003–1016, 2012.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] Tensorflow. [Online]. Available: https://www.tensorflow.org/

[19] N. Bisagno, N. Conci, and B. Zhang, "Data-driven crowd simulation," in *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*. IEEE, 2017, pp. 1–6.

[20] Z. Fang, F. Fei, Y. Fang, C. Lee, N. Xiong, L. Shu, and S. Chen, "Abnormal event detection in crowded scenes based on deep learning," *Multimedia Tools and Applications*, vol. 75, no. 22, pp. 14 617–14 639, 2016.

[21] Monitoring human activity - detection of events. [Online]. Available: http://mha.cs.umn.edu/proj_events.shtml

[22] Avenue dataset for abnormal event detection. [Online]. Available: http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html

[23] UCSD Anomaly Detection Dataset. [Online]. Available: http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

[24] UCF101 - Action Recognition Data Set, Center for Research in Computer Vision at the University of Central Florida. [Online]. Available: https://www.crcv.ucf.edu/data/UCF101.php