Violence Detection in Indoor Surveillance Cameras Using Motion Trajectory and Differential Histogram of Optical Flow

Tahereh Zarrat Ehsan
University of Guilan, Electrical Engineering Department
Iran, Guilan, Rasht
tahere.zarrat@gmail.com

Manoochehr Nahvi
University of Guilan, Electrical Engineering department
Iran, Guilan, Rasht
nahvi@guilan.ac.ir

***Abstract*—intelligent surveillance systems and automatic detection of abnormal behaviors have become a major problem in recent years due to increased security concerns. Violence behaviors have a vast diversity so that distinction between them is the most challenging problem in video-surveillance systems. In recent works, introducing unique and discriminative feature for representing violence behaviors is needed strongly. In this paper, a novel violence detection method has been proposed which is based on combination of motion trajectory and spatio-temporal features. A dense sampling has been carried out on spatio-temporal volumes along target's path to extract Differential Histogram of Optical Flow (DHOF) and standard deviation of motion trajectory features. These novel features were employed to train a Support Vector Machine (SVM) to classify video volumes into two normal and violence categories. Experimental results demonstrate that the proposed method outperforms other state-of-the-art violence detection methods and achieves 91% accuracy for detection result.**

Keywords—Intelligent Surveillance Cameras; Violence Detection; Behavior Analysis; Computer Vision

I. INTRODUCTION

Nowadays, with advances in technology and reduction of electronic system's price, the use of Closed-Circuit TeleVision (CCTV) in public or private environment is increasing dramatically. In traditional ways, a person monitors video surveillance cameras, but with the expansion of the use of these cameras in different places, monitoring with human-operator is involved errors and it may too costly. Thus, automatic video processing and abnormality detection is one of popular research issues in recent years. Most of the efforts in this field, result in the design of intelligent video surveillance cameras such that they can auto-detect of abnormal behaviors and warn security guards. The purpose of these systems is to detect abnormal behaviors among all unknown video frames in order to prevent dangerous accidents.

In analyzing surveillance videos, due to wide range of abnormal activities, detection of abnormalities is a challenging topic. Therefore, introducing a distinctive feature that detects all abnormal behaviors has become an important subject in intelligent video surveillance cameras, which several works in abnormal behavior detection are mentioned in [1, 2]. Violence detection in video sequences is a subset of the analysis of human behavior. Analysis of human behavior in the video is also used in other surveillance issues, such as falling detection in elderly homes [3], loitering detection at the airport [4] or other public places. It can also be founded on non-surveillance issues, such as movie indexing, movie rating and, human-computer interface [5].

The analysis of human behavior in order to detect abnormalities in the surveillance videos involves low and high-level steps. In the low-level step, low-level features in the videos are extracted using machine vision techniques. They are often extracted using background subtraction and target tracking methods. In these methods, background modeling is used to detect foreground target in each frame and after that moving target is tracked in the consecutive frames. These features are described with various descriptors using texture, motion, and shape information to represent the behaviors in the videos. In the high-level step, a behavior model is obtained using machine learning methods and then abnormalities can be detected. Therefore finding a suitable feature descriptor to represent human behavior is the most important part of video analysis.

In this paper, two feature descriptors based on target motion trajectory and Spatio-Temporal Interest-Points (STIP) are presented to detect violence behaviors in video sequences. Target motion trajectories indicate the global behavior but have no information of finer body movements such as hand or leg. Therefore, for such movements that all parts of the body are involved, it is essential to use local spatio-temporal feature to describe details of the motions. Spatio-temporal volumes are obtained by using trajectories information and after that, for each volume, spatio-temporal feature descriptor is calculated. A SVM classifier is trained to distinguish abnormal from common daily-life activities. The remainder of paper is organized as follows, section 2 presents a general overview and the related works in the field of violence detection, in section 3 proposed approach with some novel features is described. Experimental results and discussions are presented in section 4, and finally section 5 presents conclusion remarks.

978-1-5386-9569-2/18/$31.00 ©2018 IEEE

II. RELATED WORKS

Violence detection methods based on feature descriptors that used for action representation can be categorized into two main categories, local and global approaches. In the first method, local features are detected in a region of the frame. For example, key points in the video are extracted using STIP detectors such as Harris3D [6], Cuboid [7] and, Hessian [8] detector. In the neighboring regions of the extracted points, features are described by various descriptors such as Histogram of Optical flow (HOF) and Histogram of Oriented Gradient (HOG) [9]. For example, in [10] interest-points are extracted using Harris3D detector and in this method the magnitude of optical flow in the points is used to detect violence in videos. This feature is obtained at STIPs and sorted in each frame. The violence has occurred, if the median value of all magnitudes exceeds a specific threshold. In [11], STIPs are extracted with Harris3D but are represented by the bag of words model. In order to save computational time, among features, some of them are randomly selected and formed codebook columns. For each video, the features are represented as a function of codebook and the histogram of visual words occurrence will be obtained. In [12], a spatio-temporal volume around interest-points has been considered. DiMOLIF (Distribution of Magnitude and Orientation of Local Interest Frame) feature is calculated based on the bivariate distribution estimation for the optical flow magnitude and the orientation in the volume. In [13], a complete investigation of different STIP detectors for recognizing human interactions has been carried out. Based on their work, dense sampling approach outperforms two methods, Harris3D and Scale Invariant Feature Transform (SIFT) [14]. Disadvantages of these local methods are that, if there are too many or few motions in the videos, then feature descriptor is not discriminative.

In the global method, pixel information in the entire frame is used. In these methods, the optical flow is often used to describe motions in the video. In [15], The Optical Flow Context Histogram (OFCH) feature, which is a combination of histogram of magnitude and histogram of orientation, is obtained in the pixels of each frame. Human actions then represented by a sequence of this feature. In the proposed feature, due to large dimensions of features, Principle Component Analysis (PCA) has been used to reduce the size of them. In [16], frames are accumulated over time and the video is divided into spatio-temporal non-overlapping regions. In this method, the points are sampled densely inside the volume, and after that, optical flow is calculated. By these information, the Histogram of the Optical Flow Orientation and Magnitude and Entropy (HOFME) feature was used to describe the normal patterns in a volume of frames. HOFME is the extended of Histogram of Oriented Optical Flow (HOOF) [17] but it also uses magnitude information. In [18], dense trajectories with HOG, HOF and MBH (Motion Boundary Histograms) feature descriptor are used to detect human behaviors in videos. But in this method, feature descriptor is obtained for whole frame and in the case of multiple targets in the scene, it includes features of targets together and it is not discriminative. In [19], Violent Flows (ViF) feature has been introduced to detect abnormalities in crowded environments. The proposed feature only considers changes in the magnitude of optical flow in pixels of each frame, so abnormality in direction is not detectable. Global methods, due to the fact that they contain the whole frame information, incorporate irrelevant data into feature vector which is not related to moving targets, such as background motions and noise. These methods also cause a high computational cost because they consider entire frame in their calculations.

III. PROPOSED VIOLENCE DETECTION APPROACH

In this section, it is shown that combining DHOF and standard deviation of motion trajectory features can improve violence detection in video sequences. This method has six basic steps. Tracking of moving targets in sequential frames, calculating standard deviation of trajectory points, considering spatio-temporal volume for each target, computing optical flow, calculating DHOF feature descriptor and, training a SVM to classify video volumes. Trajectory of moving target in sequential frames is obtained using Kalman filter. Standard deviation of trajectory feature then calculated by computing the direction of trajectory points in consecutive frames. By stacking bounding boxes of moving target in sequential frames, one spatio-temporal volume for each target is extracted. Since extraction of optical flow for all pixels inside the volume is time-consuming, dense sampling method is used to extract STIPs inside the volume. The optical flow is then computed at extracted points using the Lucas-Kanade method [20]. The DHOF feature descriptor captures information based on optical flow orientation and magnitude inside the volume. In order to classify spatio-temporal volumes in the video, a SVM classifier is employed. In the training phase, it creates a hyperplane using predefined data, and in test phase, each video volume will be classify in one of the normal or violence behavior categories. The diagram of the proposed method has been shown in Figure 1 that at the first, the trajectory of the target and spatio-temporal volume is obtained, after that proposed features are calculated for each volume, and finally, a SVM is used for classification of the behaviors.

A. Tracking Estimation Using Kalman Filter

In order to estimate the movement of targets in the videos, an effective tracking method is required. There are several tracking methods reported in literatures, for instance [21,22], Among all the multi-object tracking methods, in terms of required speed and accuracy the Kalman filter [23] has appropriate performance for our purpose. Kalman filter is a recursive estimator that estimates the state of a dynamic system using a set of measurements over time. This filter generally has two steps, time update (prediction) and measurement update (correction). In the first step, the state of each target is estimated based on the state equation. In the next step, the predicted state is corrected using new measurement. In the proposed approach, for each moving target, a Kalman filter is considered. In the case of multiple targets in the scene, data association must be performed to provide a correct measurement for each moving target in the scene. The expressions for Kalman filter are [23]:

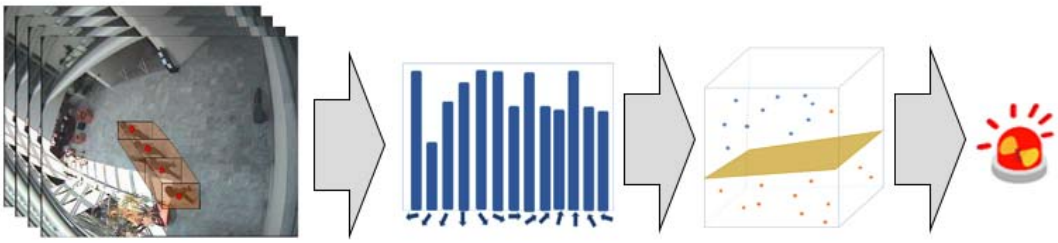

Fig 1. diagram of proposed method steps for violence detection

$$X_k = AX_{k-1} + w_k \quad , \quad z_k = HX_k + v_k \tag{1}$$

In which, A is transition matrix, w_k Gaussian process noise, H measurement matrix, v_k Gaussian measurement noise. In each frame, the state $X = (x, y)$ is extracted where x and y represent the central coordinates of the target. The target trajectory consists of state points and indicates the global movement of the target throughout the frames. In order to extract the feature descriptor, one frame is not enough to represent the actions, so the bounding box of target among the frames are used to build a spatio-temporal volume.

B. *Standard Deviation of Trajectory*

In this section, a standard deviation of trajectory feature is used to detect abnormalities in the direction of target's movement. When a person is moving normally, the angle of target's movement in short trajectory can be modeled with a normal distribution. Using this, the standard deviation of angles in short trajectory is computed to determine amount of dispersion in the angles of each target. A large standard deviation indicates abnormality in direction of movement. If a sampling rate of 25frames per second is taken into account, the time needed for frame extraction is 0.04 seconds. Due to noise in target detection, there is a probability of miscalculation in the direction of two consecutive points. Therefore, to calculate the correct angle of target's movement, two points with half-frame rate distance have been considered. So if there is noise in target detection, the angle of the target's movement will be calculated correctly. For each volume, the angles of trajectory points are obtained as follows:

$$\theta_k = \tan^{-1}\left(\frac{y_t - y_{t-0.5T}}{x_t - x_{t-0.5T}}\right), \qquad k = 1, \dots, N \tag{2}$$

Where (x_t, y_t) is the coordinate of point in frame t, $(x_{t-0.5T}, y_{t-0.5T})$ is the coordinate of point in frame $t - 0.5T$, T is video frame rate, N is number of frames in the volume. The standard deviation of motion trajectory for each volume is obtained as follows:

$$\mu = \frac{1}{N}\sum_{k=1}^{N}\theta_k \quad , \; \sigma = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}|\theta_k - \mu|} \tag{3}$$

Where μ is mean of angles and σ is standard deviation of angles. The amount of dispersion in the angles of trajectory points can be easily calculated using the above relations. A small standard deviation indicates that angles are clustered closely around the mean and a large standard deviation indicates that they are spread far from the mean and an abnormality may have occurred. So as a result, the abnormality in global movement of target is detectable.

C. *DHOF in Consecutive Volumes*

To detect violent behavior in surveillance cameras, in addition to analyzing global movement of targets, analysis of finer motion of body parts is also required. Violent behavior does not have an identical model because it includes a wide range of actions, therefore to detect violent actions accurately, it is necessary to provide a feature descriptor that can distinguish all types of them from normal behaviors. To this end, we proposed DHOF in consecutive volumes to describe variations in finer motion of target. When a person fights, his pattern of movement is constantly changing with no specific model. For example, violence behavior is composed of a series of actions such as kicking, wrestling, pushing and boxing. Therefore, for violence behavior, the histogram of optical flow in consecutive volumes does not follow a particular pattern. On the opposite, a person with normal behavior such as walking or running, has a slight difference in the histogram of optical flow in consecutive volumes. Therefore, we use DHOF along the target's path to describe its pattern of motion. To save processing time, instead of taking the total pixels into account, dense sampling of pixels in each frame is used. Each pixel is separated from its neighbors by a certain step. The number of points sampled in each frame varies according to the size of the bounding box. Dense sampling in regular locations compared to other methods of extracting points such as Harris3D, ensures that the feature points equally cover all spatial locations. The proposed feature uses magnitude and orientation information of optical flow in each of the spatio-temporal sampled points to describe the motion pattern of the volume. The magnitude and orientation of optical flow are calculated as follows [24]:

$$m = \sqrt{u^2 + v^2} \quad , \quad \theta = tan^{-1}\frac{u}{v} \tag{4}$$

Where u and v are velocity values along x and y axis. For each volume, the histogram of optical flow $H = [h_1 h_2 \dots h_B]$ is calculated. The histogram bins contain 360 degrees from -180 to 180 as follows:

$$-180 + (b-1)\frac{360}{B} < \theta < -180 + b\frac{360}{B} \tag{5}$$

Where, B is the number of bin. Each pixel in the volume, based on its orientation θ, belongs to the histogram bin b, and contributes by its magnitude m, to the sum in b. Then the obtained histograms are normalized to become scale-invariant. After that, the normalized histograms of consecutive volumes

are subtracted from each other and, by sorting the histogram bins in ascending order according to their values, DHOF is calculated. As it can be seen there are some differences between our method and HOOF such as range of orientation bins which is 180 degrees in HOOF and 360 degrees in ours, histogram is independent of motion direction in HOOF but in ours motion direction has been considered and, in HOOF histogram is normalized but in ours is not.

In figures 2 and 3, the DHOF feature in consecutive volumes is shown for two violent and normal behaviors. It has to be noted that for simplicity in this figure, 4 frames of one volume have been shown instead of all 26 frames in two consecutive volumes. Figure 2, shows two people are moving toward each other. Also their differential histograms in consecutive volumes have been shown in this figure. On the contrary, in figure 3, a violent behavior of targets with its differential histogram have been shown. As it can be seen, the proposed feature descriptor in violent scenes has larger values compared to normal behavior, due to the fact that in normal behavior, there is not any sudden changes in the direction and speed of the body parts.

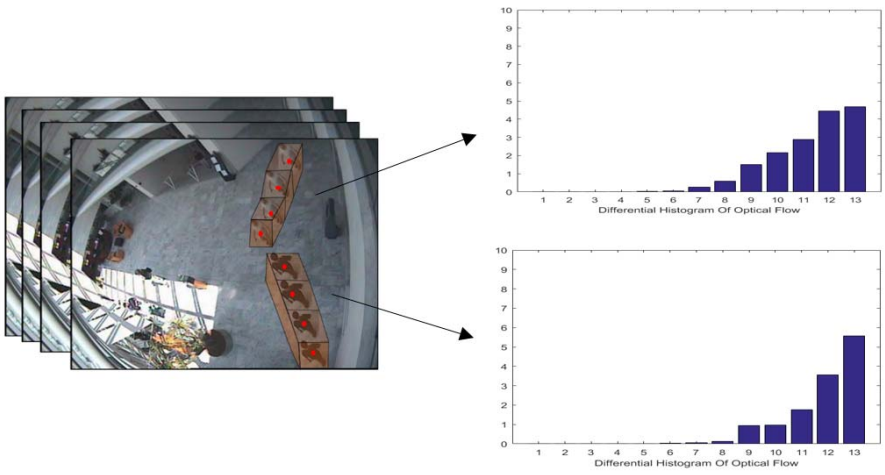

Fig 2. Normal behavior of targets in video volumes and their differential histograms

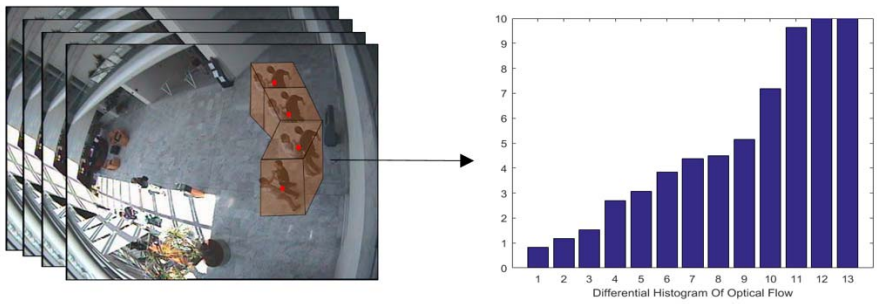

Fig 3. Violence behavior of targets in video volumes and their differential histogram

Fig 4. Detection results for two ACC and AUC metrics of proposed and other approaches

Table 1. Performance improvement of the proposed approach in comparison with others

Improvement	ACC	AUC
BoW + HOF [11]	**33%**	**35%**
BoW + HOG [11]	**31%**	**30%**
BoW + HNF [11]	**34%**	**32%**
ViF [19]	**8%**	**6%**

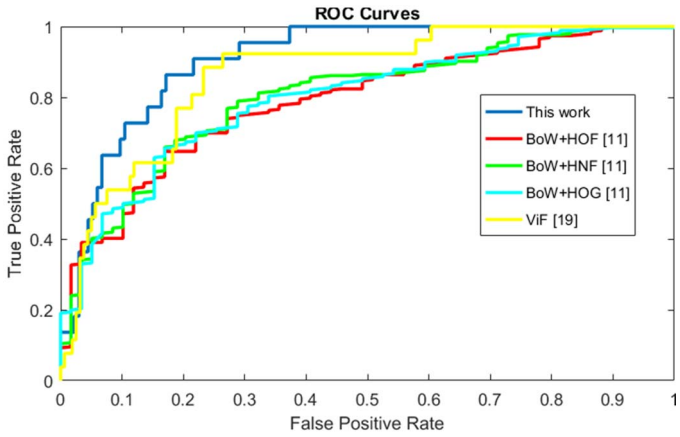

Fig 5. ROC curve for the proposed and others violence detection approaches

IV. EXPERIMENTAL RESULTS

In this section, experiments are provided to evaluate the proposed feature descriptors. Tests are performed on CAVIAR dataset [25] which included normal and violence scenes and have been taken in a building lobby. Video sequences is taken with a fixed camera with a resolution of 384x288 pixels at a rate of 25 frames per second. The proposed method has been tested with combination of DHOF and standard deviation of trajectory features. The number of histogram bins and the length of spatio-temporal volumes are set to 13. To classify video volumes into two classes of violence and normal, a 5-fold linear SVM is used. In the first step, SVM is trained with predetermined data and hyperplane is produced. In the next step, the classifier categorizes new data into the corresponding class using the hyperplane created in the previous step. In order to evaluate the proposed feature, ACCuracy (ACC) and Area Under the ROC Curve (AUC) have been adopted. These measurements have been used in most related works to evaluate different methods. The proposed method is compared with bag of words method with HOF, HNF and, HOG descriptors in [11] and Violent Flows (ViF) method in [19]. The codebook size is fixed to 500 in experiments. Figure 4 Shows ACC and AUC values for the proposed method and other approaches. As it can be seen our proposed features significantly improved the performance of detection. The ACC of the proposed method is 91% and the AUC is 93%. Table 1, shows the amount of improvement for the proposed method in comparison with others. Also the ROC curve of these methods is shown in figure 5, which, as can be seen, proposed features detect violence behavior in video better than HOF, HNF, HOG and ViF descriptors. Experiments performed on a computer with a 2.6GHz Corei7 CPU and 8.00 GB Ram. The time taken for processing each video frame is 49ms.

V. CONCLUSION

Violence detection is one of the most important issues in the computer vision techniques. In this paper, we presented the combination of DHOF and standard deviation of motion trajectory features, describing both global and local movements, to detect violent behavior in video sequences. These features are then employed for training a SVM classifier to categorize video volumes into normal and violence classes. We used dispersion and variations in the pattern of movement to accurately detect violence behaviors in videos. The experiments show that proposed feature descriptors provide better performance in comparison with methods which not used variations in magnitude and orientation of optical flow. Experimental results demonstrate the accuracy rate of 91% for violence detection by our proposed method.

REFERENCES

[1] Mabrouk, A.B. and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review" *Expert Systems with Applications*, 2017.

[2] Popoola, O. P., & Wang, K. "Video-based abnormal human behavior recognition—A review" *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 46, pp. 865-878, 2012.

[3] Sehairi, K., F. Chouireb, and J. Meunier, "Elderly Fall Detection System Based on Multiple Shape Features and Motion Analysis" *IEEE International Conference on Intelligent Systems and Computer Vision (ISCV),* pp. 1-8, 2018.

[4] Kim, Y. and Y.-S. Kim, "Optimizing Neural Network to Develop Loitering Detection Scheme for Intelligent Video Surveillance Systems" *International Journal of Artificial Intelligence,* Vol. 15, pp. 30-39, 2017.

[5] Jiang, J., Wang, Y., Zhang, L., Wu, D., Li, M., Xie, T., & Wang, S., "A cognitive reliability model research for complex digital human-computer interface of industrial system*"* *Safety Science*, 2017.

[6] Laptev, I., "On space-time interest points" *International journal of computer vision*, Vol. 64, pp. 107-123, 2005.

[7] Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S., "Behavior recognition via sparse spatio-temporal features" *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillanc,* pp. 65-72, 2005.

[8] Willems, G., T. Tuytelaars, and L. Van Gool. "An efficient dense and scale-invariant spatio-temporal interest point detector" *European conference on computer vision*, pp. 650-663, 2008.

[9] Laptev ,I., et al. "Learning realistic human actions from movies. in Computer Vision and Pattern Recognition" *IEEE Conference on CVPR,* pp. 1-8, 2008.

[10] Lyu, Y. and Y. Yang. "Violence detection algorithm based on local spatio-temporal features and optical flow" *IEEE International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII)*, pp. 307-311, 2015.

[11] De Souza, F. D., Chavez, G. C., do Valle Jr, E. A., & Araújo, A. D. A., "Violence detection in video using spatio-temporal featurs" *IEEE Conference on Graphics, Patterns and Images (SIBGRAPI),* pp. 224-230, 2010.

[12] Mabrouk, A.B. and E. Zagrouba, "Spatio-temporal feature using optical flow based distribution for violence detection*" Pattern Recognition Letters*, Vol. 92, pp .62-67, 2017.

[13] Marín-Jiménez, M.J., E. Yeguas, and N.P. De La Blanca, "Exploring STIP-based models for recognizing human interactions in TV videos" *Pattern Recognition Letters*, Vol. 34, pp. 1819-1828, 2013.

[14] Lowe, D. G. "Distinctive image features from scale-invariant keypoints" *International journal of computer vision*, Vol. 60, pp. 91-110, 2004.

[15] Chen, Y., Zhang, L., Lin, B., Xu, Y., & Ren, X., "Fighting detection based on optical flow context histogram" *IEEE International Conference on Innovations in Bio-inspired Computing and Applications (IBICA),* pp. 95-98, 2011.

[16] Colque, R. V. H. M., Caetano, C., de Andrade, M. T. L., & Schwartz, W. R., "Histograms of Optical Flow Orientation and Magnitude and Entropy to Detect Anomalous Events in Videos" *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 27, pp. 673-682, 2017.

[17] Chaudhry, R., Ravichandran, A., Hager, G., & Vidal, R., "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions" *IEEE Conference on computer vision and pattern recognition (CVPR),* pp. 1932-1939, 2009.

[18] Wang, H., Kläser, A., Schmid, C. and Liu, C.L., "Action recognition by dense trajectories" *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3169-3176, 2011.

[19] Hassner, T., Y. Itcher, and O. Kliper-Gross. "Violent flows: Real-time detection of violent crowd behavior" *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp. 1-6, 2012.

[20] Lucas, B.D. and T. Kanade, *An iterative image registration technique with an application to stereo vision,* 1981.

[21] Zhang, T., Liu, S., Xu, C., Liu, B. and Yang, M.H., "Correlation particle filter for visual tracking" *IEEE Transactions on Image Processing*, Vol. 27, pp.2676-2687, 2018.

[22] Shi, J., 1994, "Good features to track" *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Proceedings CVPR'94, pp. 593-600, 1994.

[23] Li, X., et al. "A multiple object tracking method using Kalman filter" *IEEE International Conference on Information and Automation (ICIA),* pp. 1862-1866, 2010.

[24] Bouguet, J.-Y., "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm" *Intel Corporation*, pp. 1-10, 2001.

[25] CAVIAR: Context Aware Vision using Image-based Active Recognition. Available from: http://homepages.inf.ed.ac.uk/rbf/CAVIAR./