Abnormal Behavior Detection in Crowded Scenes Using Density Heatmaps and Optical Flow

Lazaros Lazaridis, Anastasios Dimou, Petros Daras
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
Email: lazlazari, dimou, daras@iti.gr

***Abstract*—Crowd behavior analysis is an arduous task due to scale, light and crowd density variations. This paper aims to develop a new method that can precisely detect and classify abnormal behavior in dense crowds. A two-stream network is proposed that uses crowd density heat-maps and optical flow information to classify abnormal events. Work on this network has highlighted the lack of large scale relevant datasets due to the fact that dealing and annotating such kind of data is a highly time consuming and demanding task. Therefore, a new synthetic dataset has been created using the Grand Theft Auto V engine which offers highly detailed simulated crowd abnormal behaviors.**

I. Introduction

Behavior analysis is one of the most challenging tasks in computer vision. While the analysis of human activity has received a lot of attention for actions performed by individuals, work on crowded scenes has been significantly less. Crowd behavior analysis can have a large impact on a series of new application domains, such as public safety in large scale events, disaster resilience planning and transportation. Monitoring a crowd for safety and surveillance applications is essential in those areas. Automatic detection of incidents or chaotic acts in crowds and localization of the abnormal regions is invaluable to the relevant authorities.

Crowd scene analysis faces even more challenges than individual human activity due to numerous facts. The density of people found in such scenes is often prohibitive for detection algorithms that cannot identify accurately individual entities. It is even more difficult to identify body parts and their respective motion patterns to classify the individual activity of each participant. The behavior of the crowd often exhibits emergent behaviors and self-organizing activities, especially during abnormal events. Furthermore, the available content is often of low quality and it is lacking real-world examples of the events to be detected as they are only available to the authorities for legal and privacy reasons.

In this paper, a new methodology is presented to detect abnormal events in crowded scenes combining information on the density of the crowd and their motion patterns. For this purpose, a novel two-stream neural network architecture has been developed combining crowd density maps and optical flow information to identify abnormal events. In order to train the network, a new synthetic dataset has been created containing scenes with abnormal behavior. The dataset has been created using the Grand Theft Auto V (GTA V) engine and will be shared with the research community to promote work on the subject.

The contribution of this work is two-fold; (i) a novel methodology for the classification of abnormal crowd behavior is presented using crowd density heat-maps as an attention mechanism for analyzing motion patterns in crowded areas of the scene, and (ii) a new dataset is presented to allow the use of deep learning techniques on this important research area.

II. Related Work

A popular approach for abnormal event detection, due to the lack of abnormal training data, is to first learn the normal patterns, and then detect anomalies as events deviating from the normal patterns [1]. The majority of the work on anomaly detection relies on the extraction of local features from videos, that are then used to train a normality model. Trajectories have been a popular feature, detecting statistically significant deviations from the normal class to identify an anomaly [2]. However, tracking is impractical for detecting abnormal events in a crowded scene. Spatiotemporal anomalies of local low-level visual features, such as the histogram of oriented gradients [3], the histogram of oriented flows [4] and optical flow [5], by employing spatiotemporal video volumes (dense sampling or interest point selection) [6] have been also proposed. However, these approaches are bound to the quality and the completeness of the training set.

Deep learning methods have been also employed in anomaly detection. Unlike classic vision methods, optimal features are learned from the dataset. In [7], a 3D ConvNet was applied on classifying anomalies, whereas in [8] an end-to-end convolutional autoencoder was employed to detect anomalies in surveillance videos with good results. However, operations are performed only spatially, even though multiple frames are fed as input, because the 2D convolutions collapse temporal information [9]. On the other hand, Long Short Term Memory (LSTM) models are better suited for learning temporal patterns and predicting time series data. In [10], convolutional LSTMs have been proposed to learn the regular temporal sequences in videos.

In the presented work, a two-stream architecture is being proposed that employs crowd density heat-maps and optical flow respectively to detect abnormal events. Each modality is

ISBN 978-90-827970-1-5 © EURASIP 2018

fed to a network with convolutional LSTM layers to model the spatiotemporal patterns of the input. The network is trained to detect $Panic$ and $Fight$ events. A synthetic dataset has been created using the GTA V engine to train the proposed network.

III. Abnormal Event Detection

Abnormal event detection in crowded scenes is based on the analysis of the combined actions of the participating people. Due to the inapplicability of classic detection and tracking methods in highly crowded scenes, a more holistic approach is required. It is obvious that the motion content is the main source of information, either in pixel or feature level. Moreover, this motion content must be analyzed in a broader spatial and temporal context. Therefore, the optical flow of such a scene is suitable for analysis. However, such an analysis could be susceptible to errors due to similar motion content from parts of the scene where no people are present. It is argued that a crowd density heat-map can act as a driving feature to ensure that only relevant regions are included in the motion analysis. Furthermore, in the temporal domain, the changes in crowd density, such as a sudden evacuation of a place, can also provide invaluable evidence for the existence of an abnormal event.

In the following sections, the methodology used to produce the crowd density heat-maps, the extraction of the related optical flow, and the complete system architecture is presented.

A. Density Heat-Map Generation

Dense crowd images usually include a big variety of persons' head sizes due to perspective distortion. Thus, simple convolutional neural networks are patently absurd to capture characteristics of crowd at various scales. Motivated by the success of [11] we create an inception-like network, equipped with filters of different sizes in order to produce density heat-maps. Fig. 2 offers an overview of the proposed network. Those maps accurately correspond to persons' heads of different scales and angles.

The density of each annotated person head has been modeled with a delta function $\delta(x - x_i)$. This function has been convolved with a Gaussian kernel [12] so as to be converted into a continuous distribution. As a result, a density heat-map can be represented as $F(x) = H(x) * G_\sigma(x)$. Considering that the crowd in every frame is evenly distributed, the average distance $\bar{d}_i$ for each x_i annotated person head and its nearest 10 neighbor annotations can acceptably estimate the geometric distortion caused by the perspective effect using the Eq. (1). Fig. 1 illustrates a predicted density heat-map of an example image in our dataset.

$$F(x) = \sum_{i=1}^{M} \delta(x - x_i) * G_{\sigma i}, \ with \ \sigma_i = \beta \bar{d}_i \qquad (1)$$

where M is the total count of head annotations in the given image and $\beta = 0.3$, which is empirically set according to [12].

The network has been trained with large scale crowd datasets Shanghaitech Dataset [12] and UCF_CC_50 [13].

B. Optical Flow

Optical flow algorithms calculate the displacement of apparent motion of objects, surfaces and edges in a visual scene from one frame to another. Recent advances of deep learning, have enabled novel optical flow extraction techniques with impressive results. Evaluating state-of-the-art optical flow estimation algorithms such as [14] and [15] and considering accuracy as the most important factor, the Flownet 2.0 [16] was selected, which has shown remarkable results and surpasses many methods in terms of accuracy. In this work, Flownet 2.0 has been used to obtain the optical flow estimation.

C. Network Architecture

1) Preprocessing: The first step in the processing pipeline is to convert the data in a suitable format. The main task of the preprocessing stage is to convert the raw video data into a homogenized input for the model. Frames, as still images, are extracted from the raw videos and resized to 224x224 pixels. For each frame, a density heat-map and an optical flow estimation is generated. The density heat-maps are derived from the network described in Section III-A while the flow estimations are produced as described in Section III-B. Assuming that the initial extracted frames are in total N, N density heat-maps and $N - 1$ flow estimations are produced.

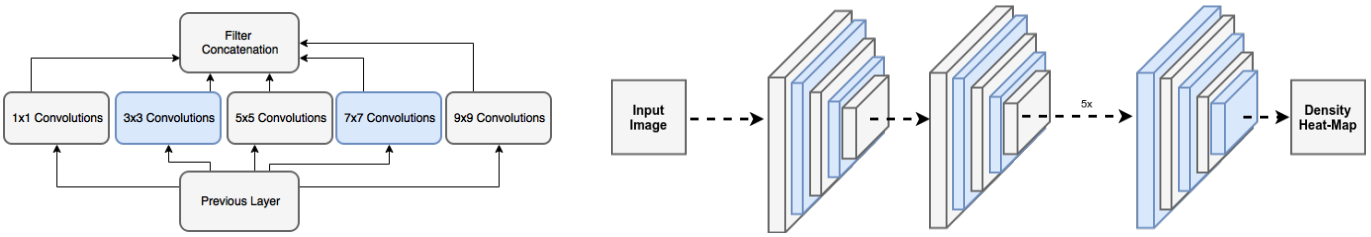
In order to train the network, consecutive batches of 10 frames are fed to the network. Training batches are collected without frame overlapping.

(a) Original frame from GTA V (b) Predicted heat-map

Fig. 1. Example frame and the predicted density heat-map.

2) The network: The proposed network consists of two streams. The first takes as input the density heat-maps (single channel $224x224$) and the second one the optical flow (two channels: distance and angle $2x224x224$) compared to the previous frame. Inspired by [17], we use a convolutional spatiotemporal network to learn the regular patterns in the training videos, in time and space. In order for the network to learn the regular patterns in the training videos, we used Long Short Term Memory networks. LSTMs are explicitly designed to avoid the long-term dependency problem. They are emplyed to acquire information about spatial structures of each input frame and for learning temporal patterns of the encoded spatial structures respectively. The outputs of each model ended up in two fully connected layers and then merged.

ISBN 978-90-827970-1-5 © EURASIP 2018



(a) Inception block (b) Density heat-map network

Fig. 2. Network architecture of the depth heat-map generator.

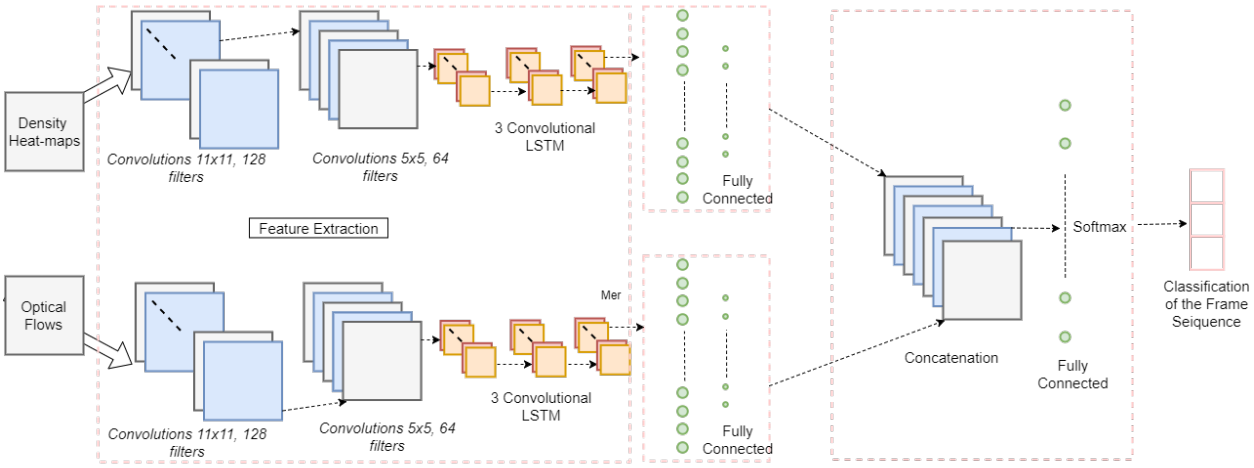


Fig. 3. Proposed network architecture for detection of abnormal events in crowded scenes.

Fig.3 illustrates the proposed architecture. Detailed parameters for each streams are listed in Table I.

Compared to [17], the convolutional spatiotemporal network is applied to two different modalities. Moreover, the network is trained with both normal and abnormal events to model their distinctive characteristics. Finally, the proposed network is designed to classify the abnormal events into more fine-tuned categories (such as $Panic$, $Fight$) rather than just flagging videos for abnormal activities.

TABLE I
DETAILED PARAMETER SETTING OF THE PROPOSED NETWORK

Layer	Parameters
Convolutional 1	11x11, 128 filters, stride 4
Convolutional 2	5x5, 64 filters, stride 2
Conv. LSTM 1	3x3, 64 filters
Conv. LSTM 2	3x3, 32 filters
Conv. LSTM 3	3x3, 64 filters
Fully Connected	50 Units
Fully Connected	3 Units
Concatenation	
Fully Connected	3 Units

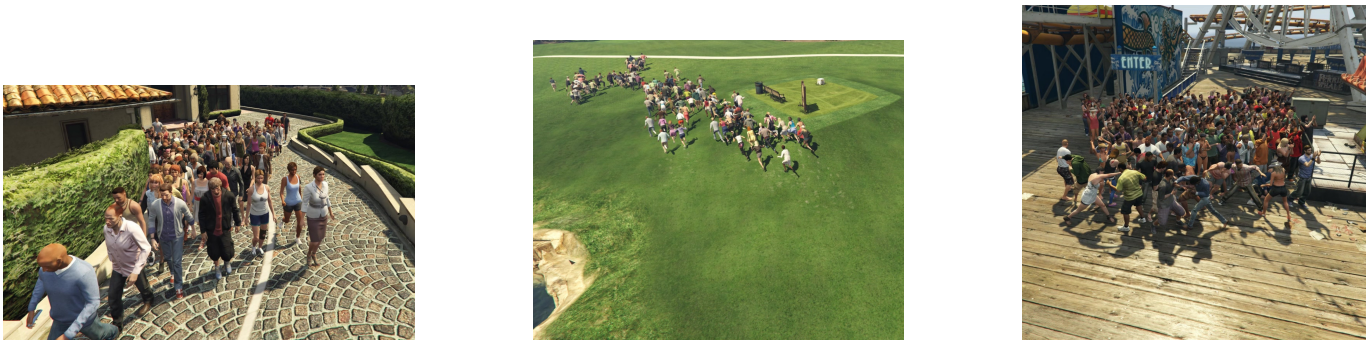
IV. DATASET

The success of training deep learning networks is dependent on the existence of large datasets. Building large, annotated with ground truth labels, datasets is extremely costly due to the amount of human time and effort needed. A good dataset for abnormal event detection in crowded scenes should be diverse, capturing all possible aspects of the problem, and precisely annotated. It should cover a variety of weather conditions, lightning conditions, different camera angles and crowd densities. Moreover, it is extremely hard to collect real data with abnormal behaviors in dense crowds due to legal and privacy regulations.

After analysing some of the most popular datasets relating to abnormal crowd behaviour, we conclude the following: (1) Violent-flows [18]: Although this dataset consists of violent crowd behavior, the average length of its videos is only 3 seconds. (2) UMN [19]: This dataset includes $Panic$ crowd behaviors, in a non-realistic way. (3) Novel violent [20]: This dataset is more realistic comparing to UMN but it lacks crowd density. We consider the last one as the most suitable to reflect the human abnormal behavior in real crowded conditions. This dataset consists of 31 video sequences. The videos were recorded as 30 frames per second. Although it lacks crowd density, each scenario is sketched in harmony with circumstances usually met in crowding issues. Last but not least, $Fight$ and $Panic$ behaviors are included in 17 out of 31 videos, among others. Some sample frames of this dataset, labeled with behavior types, are presented in Fig.5.

In order to overcome the aforementioned problems, a virtual game environment has been used. In particular, Grand Theft

ISBN 978-90-827970-1-5 © EURASIP 2018



(a) Normal Crowd (b) Crowd in Panic (c) Crowd in Fight

Fig. 4. Example frames from the abnormal crowd detection dataset. For panic and fight the frames have been cropped for visibility reasons.

Auto V was used. This game provides an extremely rich environment from which a rich dataset can be derived. It has a game engine equipped with 1.167 different pedestrian models and animals, 262 different vehicles with 14 weather and 24 hours lighting conditions in a photo-realistic projection of the real world. Examples can be seen in Fig 4.

Our dataset includes 14 videos in resolution $1920x1080$ with a framerate of 60fps, from which 7 contain normal activities and the rest contain segments with abnormal activity ($Fight$ and $Panic$). Each one is three to five minutes long, annotated with ground truth labels. Seven of them have a big variety of realistic crowded scenes where the crowd has abnormal behavior. The rest of them are normal crowded scenes where people are walking, celebrating or amusing themselves in the beach and the park. In general, the crowd behavior is annotated in 3 classes: $Normal$: normal crowd behavior, $Fight$: fight among the crowd and $Panic$: people seem to be in panic.

V. Experiments

In this section, the procedure which has been followed during the experiments is described. The aim of those experiments is dual. First, the contribution of crowd density heat-maps to improve classification of abnormal events is examined. For this purpose, the proposed methodology is compared against previous work reported in [20] for the events of $Panic$ and $Fight$. Second, initial experimental results are provided for the new dataset described.

For the purposes of this work, the dataset from [20] was filtered to exclude videos with ground truth labels that are irrelevant to this work. The remaining 17 videos have been divided in training and testing set. More specifically, we used 14 videos as training set and the rest as testing set. For each video of the training and testing set, consecutive batches of 10 frames are fed to the network.

Table III illustrates the performance comparison between our approach and the evaluations which have been done by the authors of [20] in the $Panic$ and $Fight$ behavior categories. The proposed methodology is shown to significantly outperform previously reported results.

TABLE II
Confusion matrix of the proposed method on the [20] dataset.

Actual \ Predicted	Normal	Panic	Fight
Normal	84.8	2.7	12.3
Panic	9.6	87	3.2
Fight	56.6	0	43.3

TABLE III
Comparison of classification accuracy per class.

	Normal	Panic	Fight
DT [20]	73.6	74.8	30.4
HOT [20]	79.9	62.18	38.2
Proposed	84.8	87	43.3

More detailed results for the proposed method are presented in a confusion matrix in Table II. From the confusion matrix, it is derived that the proposed method achieved 87% accuracy in events of $Panic$. Compared to $Fight$ events, $Panic$ seems to have better results. This probably happens because $Fight$ has macroscopic motion patterns similar to those occurring in other common situations, such as gatherings and loitering.

We followed the same procedure for the GTA dataset. From its 14 video sequences, we chose 10 randomly for the training set and the remaining for the testing set. Again, batches of 10 consecutive frames have been derived from both of the sets. As it can been seen in the confusion matrix, Table IV, again the $Panic$ behavior has better results comparing to the $Fight$ event. Note that due to the density of the crowd and the variety of complex events such as celebrating and amusement, this dataset seems to achieve lower overall accuracy. It is worth noting that the classification accuracy per class for the GTA V dataset, is 83.8%, 61.2% and 28.9% in $Normal$, $Panic$ and $Fight$ events respectively.

VI. Conclusion

This paper proposed a new method for the detection and classification of abnormal behavior in dense crowds. A two-stream network is proposed that uses crowd density heat-maps and optical flow information to detect $Panic$ and $Fight$ events. The proposed network has produced improved results

ISBN 978-90-827970-1-5 © EURASIP 2018

(a) Normal Crowd (b) Crowd in Panic (c) Crowd in Fight

Fig. 5. Example frames from the dataset presented in [20].

TABLE IV
CONFUSION MATRIX OF THE PROPOSED METHODOLOGY ON THE $GTA - Crowd$ DATASET

Actual \ Predicted	Normal	Panic	Fight
Normal	83.8	5.8	10.2
Panic	37.3	61.2	0.6
Fight	70.2	0.7	28.9

compared to the ones reported in literature. While there is a clear improvement in results, the lack of large scale relevant datasets has been highlighted. Therefore, a new synthetic dataset has been created using the Grand Theft Auto V engine which offers highly detailed simulation crowd abnormal behaviors. This dataset will be made available to the community.

As a next steps for this work, both the GTA V dataset extension and the use of the synthetic dataset to improve detection accuracy, employing transfer learning principles is foreseen.

ACKNOWLEDGMENT

This work was supported by the European Project: SURVANT http://survant-project.eu/ Grant no. 720417 within the H2020 FTIPilot-2015.

REFERENCES

[1] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in Matlab," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2720–2727.

[2] S. Zhou, W. Shen, D. Zeng, and Z. Zhang, "Unusual event detection in crowded scenes by trajectory analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1300–1304.

[3] T. Xiao, C. Zhang, H. Zha, and F. Wei, "Anomaly detection via local coordinate factorization and spatio-temporal pyramid," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 66–82.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008.

[5] V. Reddy, C. Sanderson, and B. C. Lovell, "Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. IEEE, 2011.

[6] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 65–72.

[7] S. Zhou, W. Shen, D. Zeng, M. Fang, Y. Wei, and Z. Zhang, "Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes," *Signal Processing: Image Communication*, vol. 47, pp. 358–368, 2016.

[8] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 733–742.

[9] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4489–4497.

[10] J. R. Medel and A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," *arXiv preprint arXiv:1612.00390*, 2016.

[11] L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang, "Multi-scale convolutional neural networks for crowd counting," *arXiv preprint arXiv:1702.02359*, 2017.

[12] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 589–597.

[13] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2547–2554.

[14] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2017.

[15] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2758–2766.

[16] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2017.

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

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

[19] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.

[20] H. Rabiee, J. Haddadnia, H. Mousavi, M. Kalantarzadeh, M. Nabi, and V. Murino, "Novel dataset for fine-grained abnormal behavior understanding in crowd," in *Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*. IEEE, 2016.

ISBN 978-90-827970-1-5 © EURASIP 2018