Journal of Advanced Research in Technology and Management Sciences
Volume: 04 Issue: 04 ISSN: 2582-3078 August-December 2022
Available online at: http://www.jartms.org

Anomaly Detection Using Transfer Learning in CNNs

[1]KALLEPALLI ROHIT KUMAR, [2]DR.NISARG GANDHEWAR

Department of Computer Science and Engineering
Dr A.P.J. Abdul kalam University, Indore-452010(INDIA)

Abstract

The demand for greater security measures in crowded environments for monitoring and protecting operations has made video anomaly detection significant study area in computer vision. This is due to the fact that the detection of anomalies in video surveillance systems has increased, making it one of the leading focus areas in the current field of research. In this study, we offer a method for identifying out-of-the-ordinary anomaly behaviour in video footage of crowded settings. The proposed approach uses the three CNN architectures of AlexNet, ResNet and VGGNet. Fine tunes the architectures by using conjugate gradient optimization. The transfer learning approach is followed for the classification. The proposed method saves time by not performing the training part again and combines the results from the three architectures. Random forest and Softmax classifier perform the classification. The proposed model faresbetter than some of the existing literatures. The study was performed on three datasets namely Ped1,Ped2 and Avenue.

Keywords: AlexNet, ResNet ,VGGNet, classifier, anomaly, Random Forest, Softmax

1 Introduction

When there is a very low probability of incidents that require follow-up, monitoring surveillance footage can be an incredibly time-consuming process. The complexity of normal crowd behaviour makes this challenge far more challenging when dealing with crowded settings. More surveillance cameras are being put up in busy public places to make them safer and keep an eye on what's going on every day.[1][2] With the rise of terrorist threats and other crimes, intelligent video monitoring has become an integral part of security systems for detecting and neutralising threats at crowded places like bus stations, railway stations, airports, stadiums, etc.

When people gather in vast numbers to form a crowd, the situation becomes far more complicated in terms of monitoring than when it involves only a few people. Counting, estimating, and observing crowds are only a few of the many possible approaches to the study of crowds. The primary aim of crowd anomaly detection is to spot out-of-the-ordinary crowd behaviour, such as a car driving down a sidewalk or a strange pattern of individuals rushing erratically after a mishap. Anomalies are defined as out-of-the-ordinary data patterns. [3][4][5] Credit card fraud prevention and email account two-step verification are two widespread applications of anomaly detection. Detecting intrusion attempts is its primary use in the networking industry. Figure 1 gives the architecture of the three CNN used for the study namely AlexNet, ResNet and VGGNet. AlexNet is an example of a convolutional neural network (CNN) architecture that was developed through a cooperative effort by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. AlexNet is a deep neural network that consists of 8 layers. AlexNets implemented a few innovative strategies, including the RelU activation function and Dropout layers. ResNet is an artificial neural network that pioneered the use of a so-called "identity shortcut link." This connection enables the model to bypass one or more levels of the neural network. Because this method is used, it is possible to train the network on hundreds of levels without negatively impacting its performance. It has quickly emerged as one of the most widely used architectures for a wide variety of computer vision tasks. The performance of a wide variety of computer vision applications, including object detection and facial identification, has been improved by taking advantage of its powerful representational ability. These applications include image categorization as well as facial

recognition and object detection. Karen Simonyan and Andrew Zisserman from Oxford proposed VGGNet in 2014. This design has 13 convolutional and 3 fully linked layers. All buried layers have ReLU nonlinearity. This design features 33 convolutional kernels. Thirteen convolutional layers can form five groups. Block 1 has 2 floors. It has 64 channels. The second block has 128 channels and two convolutional layers. The 512-channel, 3-convolutional-layer third block has 512 channels. Final two blocks have 512 channels and three convolutional layers. Each block has a 22 max-pooling layer. Once the fifth block is in place, the next three layers can be built. First two have 4096 channels. Impressive 3,000-channel tier. Fully linked layers can be adjusted for different datasets and activities.

Some of the problems that need fixing are abandoned items, illegal traffic, parking in restricted areas, fires, deaths, break-ins, and acts of violence. Can be uncovered by employing anomaly detection techniques on surveillance footage of the incidents in question. [6] Odd behaviours like jaywalking, loitering, trespassing, crawling, and killing can all be uncovered with this approach. The main objectives of the article are

1. To find anomalous behaviour in a crowded environment
2. Use transfer learning to implement the approach.

- Use three different CNN architectures AlexNet, ResNet and VGGNet for the study

1. Implement anomaly detection using fine tuning and transfer learning approach.

The introduction section is followed by a literature survey of related works (Section 2). The methodology is explained in Section 3. The experimental results and their analysis are detailed in Section 4, followed by a conclusion.

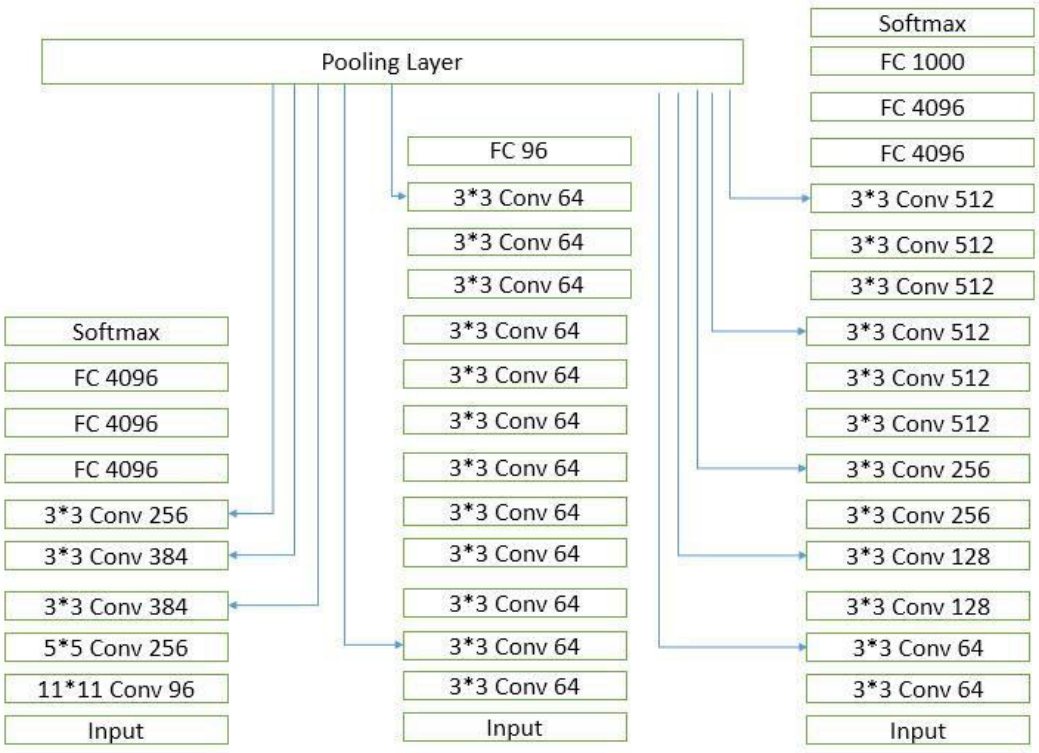

Figure 1: Architecture of AlexNet ResNet and VGGNet

Journal of Advanced Research in Technology and Management Sciences
Volume: 04 Issue: 04 ISSN: 2582-3078 August-December 2022
Available online at: http://www.jartms.org

2 Related Work

Sabokrou et al. [14] A quick and trustworthy approach to spotting and pinpointing anomalies in video footage of busy events. This work addresses the continued difficulty of timely localization. They suggest a cubic patch-based technique, distinguished by cascade of classifiers and utilising a cutting-edge feature-learning strategy. The classifier cascade consists of two distinct phases. To begin, "many" normal cubic patches are quickly identified using a lightweight yet deep 3D auto-encoder. After initially processing data in the form of small cubic patches, the network's second stage involves expanding the remaining candidates and evaluating them with a more advanced and deeper 3D convolutional neural network. Using the multi-view representation learning framework, Deepak et al. [13] proposed novel methods intwo directions. .The first method is a multi-view model learner that uses a mix of deep features found by a 3D spatiotemporal autoencoder and robust features made by hand based on the autocorrelation of gradients in space and time .Second, we use deep multi-view representation learning to detect anomalies by combining deep features extracted from two streams. Gunale and Mukherji [12] proposed a method forther extraction of spatiotemporal features from data on optical flows in order to detect outliers. The description analyses videosusingd clever surveillance technology. The training phase of deep learning is where the system absorbs bote high-level and low-level information and uses it to learn all the typical patterns. The system handles the challenge of detection under different transformations with respect to the state-of-the-art methods, and it does so robustly by identifying both local and global abnormal occurrences from complicated scenarios. Singh et al. [11] proposed anomaly identification in video footage displaying crowded scenes using pre-trained ConvNets, a set of classifiers, and the concept of Aggregation of Ensembles (AOE). The proposed approach utilises a suite of tunedconvolutional Neural Networks (CNN) on the assumption that different CNN architectures pick up on variable degrees of semantic representation from crowd recordings, allowing for enriched feature sets to be recovered. Roshtkhari et al. [15] have used probability theory for the early identification of anomalies. They are creating the codebook methodically, with a set of video phrases as their base. This approach involves incrementally modifying a likelihood density function in order to identify new baseline habits. Kim et al [16] used Markov model in random field as a means of detecting anomalies in video data. It deals with anomaly detection on two levels: locally and globally. Locally, it can distinguish out-of-the-ordinary behaviour inside a busy environment, whileglobally, it can handle aberrant interactions between activities taking place in different parts of the scene. That's right; it's the one in charge of spotting irregularities. Duque et al.[7] proposed two methods to detect abnormal behaviors,The first approach uses principal component analysis (PCA) to extract features from the blobs in order to detect anomalous activities. Using support vector machine, anomalous behaviour can be classified as normal or abnormal following feature extraction. Optical flow is another technique for spotting outliers in a busy scene.

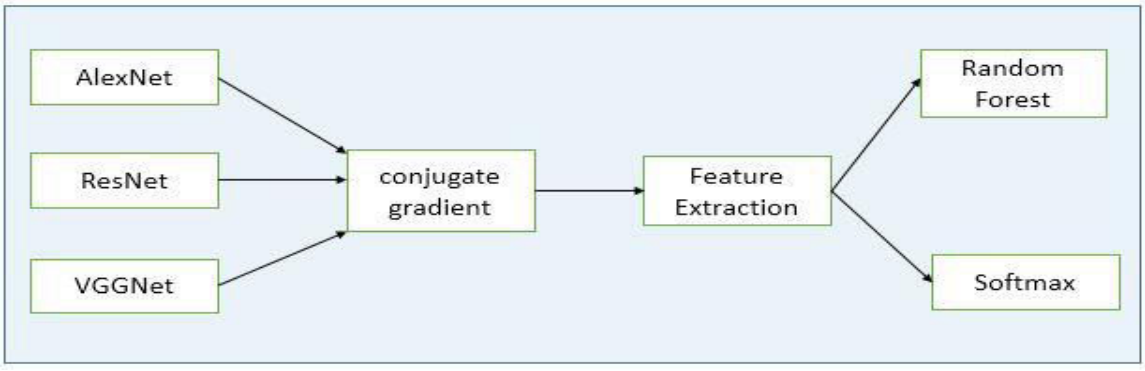

Figure 2: Process Flow diagram

3. Methodology

We used the following methodology to investigate the changing effects of various CNN features for detecting anomalies in crowded surveillance environments: Figure 2 shows the proposed model's process flow. The optimization of the features selected for the study was performed using conjugate gradients (CG). For the solution of nonlinear optimization issues and large-scale linear systems of equations, the CG method is applied. The first-order approaches' convergence rate is sluggish. The second-order approaches, however, require a lot of resources. A middle-ground strategy known as conjugate gradient optimization combines the benefits of first-order information with the rapid convergence of high-order techniques. The study performs optimization based on conjugate gradient to analyse the result of the fine tuning step before feeding the respective architectures, namely AlexNet, ResNet, and VGGNet. Before combining them all, these network architectures are fine-tuned using CG optimization to perform result analysis on individual architectures. This step helps us validate the individual performance of the three networks. Further comparison can also be made after the next phase, where the three architectures are combined to form an ensemble model. Table 2 gives the training time that the respective models take for fine tuning using the CG optimization algorithm.

4 Experimental Results Analysis

The proposed method is compared to the state-of-the-art algorithms, both qualitatively and quantitatively, and the results are discussed in this section. The implementation of the algorithm is performed by using python on DELL PowerEdge R540 Server, 32 GB RAM, Intel Xeon processor.

4.1 Dataset

The CUHK Avenue dataset was compiled using a stationary 640 x 360-pixel video camera that was capturing path movement at the City University of Hong Kong. This compilation of videos contains 16 training videos of typical human behaviour and 21 videos of behaviour that is not considered normal. These video sequences are available in the Ped-1 and Ped-2 databases as individual frames, while the Avenue dataset offers a variety of individual videos. Therefore, the Avenue dataset videos were segmented to obtain a new frame after a fixed interval of 0.5s. These three datasets simulate typical and unusual behaviour in crowded settings and consist of both normal and anomalous behaviour. Table 1 gives a description of the three datasets.

Table 1 Dataset Features

Dataset	No. of Normal Data	No. of Anomalous Data	Dimension
UCSD Ped 1	58	60	238*158
UCSD Ped 2	28	22	360*240
Avenue	42	48	640*360

70% of the respective records in each class of the dataset were retained for forming the training dataset and the rest 30% formed the test dataset. A total of 40, 19, and 29 normal video samples from the Datasets, and a consists of 42, 15, and 33 anomalous video samples were used for training, respectively.

Due to the huge size of the vector dimensions of the input video frames, the problem of overfitting will take place. To avoid the overfitting The dimensionality reduction was performed issue with the

proposed model, dimensionality reduction was performed by using Principle Component Analysis (PCA). Using orthogonal transformation,

4.2 Pre-processing

Pre-processing of the input samples is done and the dimensionality of all the samples is set to 224x224 pixels. The pixels are normalized in a range from 0 to 1. On the basis of the frame rate of training data the depth of vector is selected, which corresponds to a period of about one-third of a second.

Table 2 Training time

Model	AlexNet			ResNet			VGGNet		
Dataset	Ped1	Ped2	Avenue	Ped1	Ped2	Avenue	Ped1	Ped2	Avenue
Training time(Min)	63	58.5	53.6	74.5	71.3	93.5	97.5	87.4	94.5

For its ensemble approach, Random Forest employs bagging, whereas its individual model, the Decision Tree, provides the basis for the ensemble. After optimising the three architecture networks with fine tuning using the optimization method, The random forest and softmax classifiers were used for further study. Only the random forest classifier was considered for the transfer learning process. The proposed approach is to ensemble the output from the three architectures and perform an averaging process to check for accuracy. Table 3 contains the output of the accuracy of the three networks with the Ped 1, Ped 2, and Avenue datasets, both in individual cases and also when the proposed ensemble averaging approach is used. It was found that having a larger number of classifiers increased their accuracy more than having fewer classifiers. The ensemble approach of having a random forest classifier with a softmax classifier has better accuracy than having two separate classifiers. The VGGNet has individual accuracy better than both the AlexNet and ResNet. The proposed ensemble approach integrates the three architectures and performs an averaging of the outputs. Hence, it has better accuracy than the individual classifiers, as is evident from the accuracy value in Table 3.

We draw the ROC and calculate the area under the curve (AUC) to determine how well our algorithm performs in terms of true positives and false negatives. Figures 3 and 4 depict the ROC curves for the ped 1 and ped 2 datasets, as well as results comparisons with the standard literature.

Table 3 Accuracy performance

Ensemble	Model	UCSD Ped 1	UCSD Ped 2	Avenue
Softmax	AlexNet	83.3	81.6	80.2
	ResNet	84.2	83.8	82.4
	VGGNet	83.7	85.6	84.2
Random Forest	AlexNet	78.8	75.7	73.2
	ResNet	81.2	80.3	78.9
	VGGNet	84.6	83.8	81.2
Proposed	AlexNet	84.3	83.2	79.5
	ResNet	86.2	84.5	84.6
	VGGNet	87.8	86.4	87.8

Literature	PED 1	PED 2

Khan et al.[21]	80.08	93.71
Xu. et al.[20]	84.4	86.8
Cong et al.[18]	54.1	84.2
Revathi&Kumar[17]	67.3	71.9
Proposed	94.8	95.3

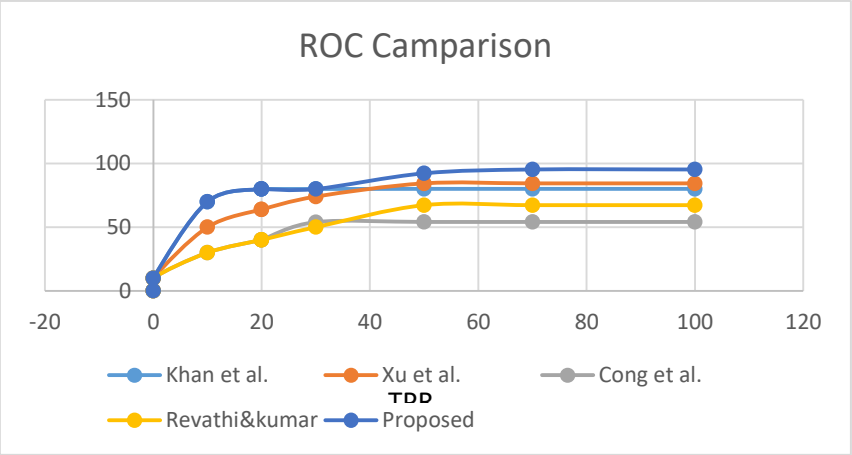

Figure 3: ROC curve for PED 1 Dataset

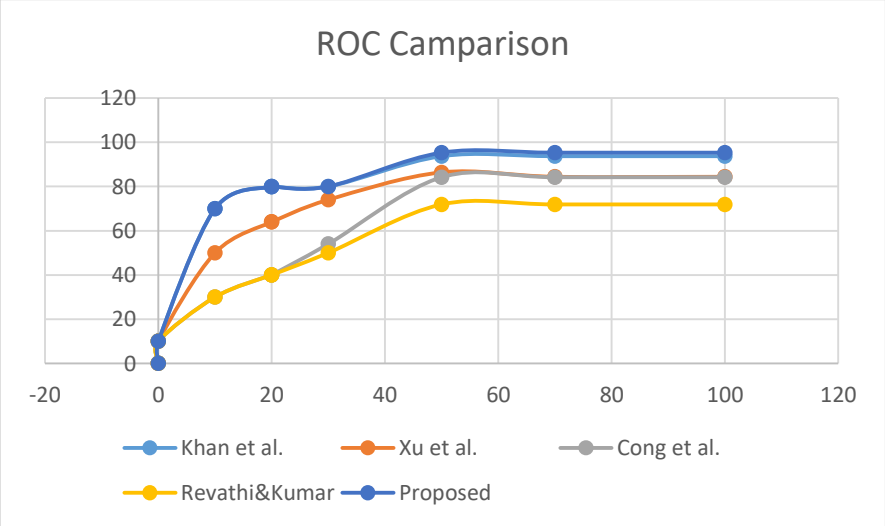

Figure 4: ROC curve for PED 2 Dataset

Even in scenarios where there is limited data, the proposed method outperforms existing state-of-the-art literatures. As our experiments have shown, there are not a lot of samples included in most crowd behaviour datasets, which makes it challenging to train CNN models with less number of data. The proposed approach performs an averaging of the outputs of the three architectures. The classification part is not again performed in the proposed approach as it depends upon the concept of transfer learning. Skipping the classification time saves a lot of time.

Conclusion

This work introduces the notion of combining the outputs of architecture models, which is a method of utilising the already-available capacity of pre-trained convolutional networks to extract high-level features that may be utilised in the succeeding steps. The article presents a model to detect anomalous behaviours in crowded places under camera surveillance. The approach uses conjugate gradient to perform the optimization of the networks, and the fine tuning of the architectures of AlexNet, ResNet, and VGGNet results in the identification of anomalous behaviour. The proposed model outperforms many of the existing literatures, and the concept can be applied to studies involving transfer learning and classifier combining.

References:

[1] Saligrama, V., Chen, Z.: Video anomaly detection based on local statistical aggregates. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2112–2119. (2012)

[2] Sabokrou, M., et al.: Deep-cascade: cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Trans. Image Process. 26(4), 1992–2004 (2017)

[3] Nair, V., Kosal Ram, P. G., & Sundararaman, S. (2019). Shadow detection and removal from images using machine learning and morphological operations. *The Journal of Engineering*, *2019*(1), 11–18. https://doi.org/10.1049/joe.2018.5241

[4] Rajesh Banala, D.Upender,: "Remote Home Security System Based on Wireless Sensor Network Using NS2", International Journal of Computer Science and Electronics Engineering, India, Vol. 2 Issue 2 (2012).

[5] Hasan, M., et al.: Learning temporal regularity in video sequences. In:2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733–742. (2016)

[6] Xu, D., et al.: Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. Neurocomputing. 143, 144–152 (2014)

[7] Duque D, Santos H, Cortez P (2007) Prediction of abnormal behaviors for intelligent video surveillance systems. In: Computational intelligence and data mining, 2007. CIDM 2007. IEEE Symposium on, pp 362–367

[8] K. K. Verma, B. M. Singh, and A. Dixit, "A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system," Int. j. inf. tecnol., vol. 14, no. 1, pp. 397–410, Feb. 2022, doi: 10.1007/s41870-019-00364-0.

[9] A. Aboah, M. Shoman, V. Mandal, S. Davami, Y. Adu-Gyamfi, and A. Sharma, "A Vision-based System for Traffic Anomaly Detection using Deep Learning and Decision Trees," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, Jun. 2021, pp. 4202–4207. doi: 10.1109/CVPRW53098.2021.00475.

[10] Weixin Li, V. Mahadevan, and N. Vasconcelos, "Anomaly Detection and Localization in Crowded Scenes," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 1, pp. 18–32, Jan. 2014, doi: 10.1109/TPAMI.2013.111.

[11] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia, "Crowd anomaly detection using Aggregation of Ensembles of fine-tuned ConvNets," Neurocomputing, vol. 371, pp. 188–198, Jan. 2020, doi: 10.1016/j.neucom.2019.08.059.

[12] K. Gunale and P. Mukherji, "Deep Learning with a Spatiotemporal Descriptor of Appearance and Motion Estimation for Video Anomaly Detection," J. Imaging, vol. 4, no. 6, p. 79, Jun. 2018, doi: 10.3390/jimaging4060079.

[13] K. Deepak, G. Srivathsan, S. Roshan, and S. Chandrakala, "Deep Multi-view Representation Learning for Video Anomaly Detection Using Spatiotemporal Autoencoders," Circuits Syst Signal Process, vol. 40, no. 3, pp. 1333–1349, Mar. 2021, doi: 10.1007/s00034-020-01522-7.

[14] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-Cascade: Cascading 3D Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes," IEEE Trans. on Image Process., vol. 26, no. 4, pp. 1992–2004, Apr. 2017, doi: 10.1109/TIP.2017.2670780.sss

[15] Roshtkhari, M.J., Levine, M.D.: An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. Comput. Vis. Image Underst. 117(10), 1436–1452 (2013)

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

[17] A.R. Revathi , D. Kumar ,An efficient system for anomaly detection using deep learning classifier, Signal Image Video Process 11 (2) (2017) 291–299 .

[18] Y. Cong , J. Yuan , Y. Tang , Video anomaly search in crowded scenes via spa- tio-temporal motion context, IEEE Trans. Inf. Forensics Secur. 8 (10) (2013) 1590–1599 .
[19] M. Sabokrou , M. Fathy , H. M. , R. Klette , Real-time anomaly detection and lo- calization in crowded scenes, in: Proceedings of the CVPRW, Boston, 2015 .
[20] D. Xu , R. Song , X. Wu , N. Li , W. Feng , H. Qian , Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts, Neuro- computing 143 (2014) 144–152 .
[21] M.U. Khan , H. Park , C. Kyung , Rejecting motion outliers for efficient crowd anomaly detection, IEEE Trsans. Inf. Forensics Secur. 14 (2) (2019) 541–556 .