HR-Crime: Human-Related Anomaly Detection in Surveillance Videos

Kayleigh Boekhoudt[1(✉)], Alina Matei[1], Maya Aghaei[2], and Estefanía Talavera[1,3]

[1] University of Groningen, Groningen, The Netherlands
k.j.boekhoudt@student.rug.nl
[2] NHL Stenden University of Applied Sciences, Leeuwarden, The Netherlands
[3] University of Twente, Enschede, The Netherlands

Abstract. The automatic detection of anomalies captured by surveillance settings is essential for speeding the otherwise laborious approach. To date, UCF-Crime is the largest available dataset for automatic visual analysis of anomalies and consists of real-world crime scenes of various categories. In this paper, we introduce *HR-Crime*, a subset of the UCF-Crime dataset suitable for *human-related* anomaly detection tasks. We rely on state-of-the-art techniques to build the feature extraction pipeline for human-related anomaly detection. Furthermore, we present the baseline anomaly detection analysis on the HR-Crime. HR-Crime as well as the developed feature extraction pipeline and the extracted features will be publicly available for further research in the field.

Keywords: Forensics · Human-related anomaly detection · Surveillance videos

1 Introduction

The detection of anomalous events in videos is a challenging task due to the broad definition of the term 'anomaly', as well as insufficient annotated data. Despite this, there has been much research in the field of video surveillance anomaly detection in the past years [15]. Surveillance cameras are a widely used technology which aids law enforcement agencies in ensuring general public safety. Surveillance footage is also considered a reliable piece of forensic evidence when the anomalies captured on the footage are identified as crimes. However, due to the overwhelming amount of surveillance video data (usually, surveillance cameras transmit 24/7 live footage), there is an outstanding need for the automation of abnormality detection in such videos.

M. Aghaei and E. Talavera—Contributed equally.

© Springer Nature Switzerland AG 2021
N. Tsapatsoulis et al. (Eds.): CAIP 2021, LNCS 13053, pp. 164–174, 2021.
https://doi.org/10.1007/978-3-030-89131-2_15

Fig. 1. Feature extraction pipeline of HR-Crime dataset. Given the frames of a video, we first extract human proposals using YOLOv3-spp [5]. Second, AlphaPose [3] is applied to detect body skeletons. Finally, PoseFlow [17] is used to track skeletons.

Human-related anomaly detection in surveillance videos, as a specific case of anomaly detection, is closely related to human activity detection that can be visually recognized as abnormal through body movement. In recent years, there have been many advances in the field of human pose (skeleton) estimation. Generally, there are two main types of frameworks used for pose detection. Two-step frameworks such as AlphaPose [3] use the top-down approach for pose detection. The idea is to first use an object detector to find people and then perform single-person pose estimation for each detected person. In contrast, methods that use the bottom-up approach to detect poses, first localize body parts and then group them into person instances [1,9].

There are advantages and disadvantages to these methods. Bottom-up methods feed the whole image to their architecture, which may impose limitations on the input image size. On the other hand, top-down methods crop and feed each individually detected human bounding boxes to their architecture. The disadvantages of the top-down method is that the keypoint detection performance depends on the quality of the bounding boxes and that the runtime is proportional to the number of people in the frame. Bottom-up approaches, on the contrary, do not have the issue of early commitment and runtime complexity. Cao et al. [1] suggests that AlphaPose [3] to be used for maximum accuracy, OpenPose [1] for maximum speed, and METU [9] for a trade-off between them.

Many efforts have also been made in recent years towards accurate human pose tracking. PoseTrack [8] and ArtTrack [6] introduced the multi-person pose tracking incorporating the idea of the part-based pose estimator DeeperCut [7] by extending spatial joint graph to spatio-temporal graph. First, the model generates for each frame, a set of detected keypoints and constructs the spatio-temporal graph. The model then solves an integer linear program to divide the graph into sub-graphs that correspond to skeleton trajectories of each person. These methods are also known as jointing schemes and are computationally heavy and not scalable to long videos. Top-down methods such as PoseFlow [17] are more scalable. The model starts by detecting human bounding boxes in every frame and extracts poses from each bounding box. The boxes are then

tracked over the entire video in terms of similarity between pairs of boxes. These types of pose trackers are therefore also called Detect-And-Track methods. PoseFlow [17] also takes motion and pose information into account by implementing a cross frame matching technique to propagate box proposals to previous and next frames.

Extending on the task of video anomaly detection, small progress has been made targeting the human-related anomaly detection plainly [2,4]. In [12], the authors proposed the MPED-RNN architecture for anomaly detection in surveillance videos based on skeleton trajectories described by local and global body movement. The proposed MPED-RNN follows an encoder-decoder architecture: the encoder learns close approximations of normal trajectory which are decoded with high accuracy; this implies that, when presented with abnormal trajectories, the encoder-decoder architecture obtains inaccurate reconstructions which results in high anomaly scores.

One reason for the small progress in human-related anomaly analysis might be the lack of human centered anomaly related datasets. Hence, the main contributions of this work are planned to target this shortage as following:

1. We introduce and make publicly available the Human-Related Crime dataset (HR-Crime) together with the annotations at the frame level.[1]
2. We present baseline results on HR-Crime intending to contribute to future research in the field of human-related anomaly detection.

The rest of the paper is organized as follows. In Sect. 2, we discuss our feature extraction pipeline for development of the HR-Crime dataset. In Sect. 3, we elaborate on the implementation details and discuss the obtained results. Finally, we draw conclusions in Sect. 4.

2 HR-Crime Dataset

As mentioned earlier, Morais et al. recently introduced the only work on human-related anomaly detection in surveillance videos [12]. Their introduced architecture, MPED-RNN, requires a defined set of features extracted from videos to detect the human-related abnormalities. In an attempt to provide the baseline results on HR-Crime, we opt for extracting the required features from the UCF-Crime [16] videos and only keep the relevant information out of it to build the HR-Crime dataset. In this section, we describe the followed steps to prepare the HR-Crime dataset.

[1] Dataset is publicly available at https://doi.org/10.34894/IRRDJE.

2.1 HR-Crime Statistics

The UCF-Crime dataset [16] consists of 950 real-world surveillance videos of anomalies, and 950 normal videos. The anomalies are divided into 13 categories: *Abuse*, *Arrest*, *Arson*, *Assault*, *Burglary*, *Explosion*, *Fighting*, *Road Accidents*, *Robbery*, *Shooting*, *Shoplifting*, *Stealing* and *Vandalism*. Duplicates may occur because some videos either have multiple anomalies or the anomaly may fall into more than one category. UCF-Crime dataset as is originally gathered to represent the *in-the-wild* nature of the crime scenes, at times lacks the required clarity in content even for human eyes. This comes in addition to the fact that only a subset of it is human-related. Hence, for further human-related anomaly analysis, we extracted HR-Crime out of UCF-Crime dataset using the following guidelines:

- Omitting videos of anomalous events that are not human-related. We refer to 'human-related' if the main performing subjects are human. Within this definition, dog abuse is not considered human-related.
- Excluding videos that do not have a clear view of the people at the scene.
- Ignoring videos with large crowds, as our goal is not to do crowd analysis which is essentially a different task than human behavior analysis.
- Ignoring videos longer than 100 min.

The resulted HR-Crime dataset consists of 789 human-related anomaly videos and 782 human-related normal videos. Examples are shown in Fig. 2. Table 1 shortly describes UCF-Crime and the newly proposed HR-Crime datasets. HR-Crime consists of 239 testing videos with annotation. Each video frame in this test set is annotated as normal or anomalous. As it can be observed, most categories consist mainly out of human-related videos. For instance, for *Shoplifting*, all the videos are human-related. In contrast, *Road accidents* has relatively the least number of human-related videos, which is expected as the people are mostly in cars, hence, not visible in the cameras.

Figure 3a compares the range and distribution of the video length in minutes for different categories. We observed that the video length varies for all categories. *RoadAccidents* has the smallest variability in length compared to the other categories. In contrast, videos from the categories *Shoplifting*, *Arrest*, *Fighting* and *Burglary* vary the most in length. However, the *Normal* category has the most number of videos that are longer than other videos of the same category. Inspecting the HR-Crime dataset, we realized that the longest *Normal* video lasts 93.62 min.

A comparison of the range and distribution of the number of tracked skeletons is shown in Fig. 3b. *Arson* varies the least in number of detected people. In contrast, *Arrest*, *Normal*, *Shoplifting* and *Fighting* range the most. After further analysis we realized that the *Normal* category has a video with a maximum of 1084 skeleton trajectories.


(a) Abuse

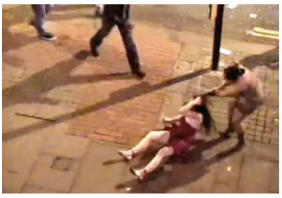

(b) Arrest

(c) Arson


(d) Assault

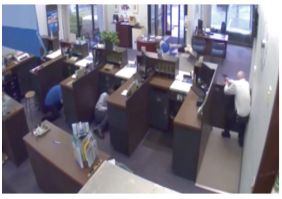
(e) Burglary


(f) Explosion


(g) Fighting


(h) Road Accidents


(i) Robbery


(j) Shooting


(k) Shoplifting


(l) Stealing


(m) Vandalism

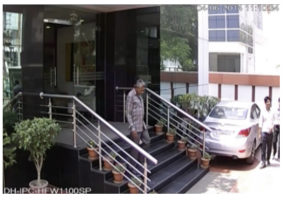
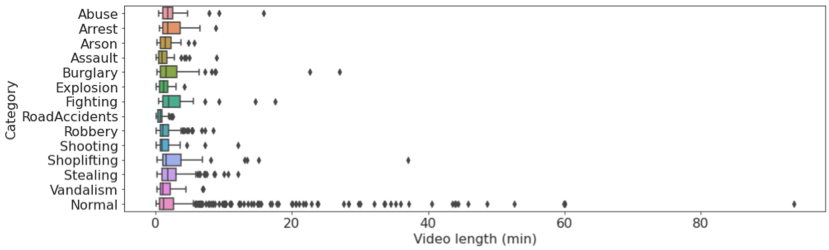
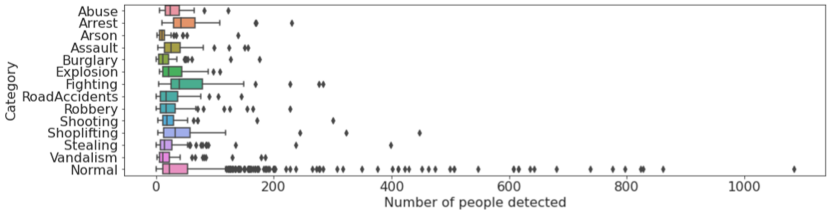

(n) Normal

Fig. 2. From (a) to (n), example frames from HR-Crime: a man punches an older woman; the police escort a man; a person pours inflammable liquid on the ground; a man holds another man by his neck; two burglars steal an ATM; people run away from an explosion; a woman drags another woman by her hair; a car hits a child on a bicycle; a robber holds a knife; a shooting in an office; a man shoplifts a watch; a man hot-wires a motorcycle; a vandal writes on a garage door; a man walks down some stairs.



(a) Video length in minutes



(b) Skeleton trajectories

Fig. 3. Video length and number of skeleton trajectories distribution in HR-Crime.

Table 1. The first three rows display the number of videos, HR-videos and HR-test videos per category in UCF-Crime dataset, respectively. The number of anomalous and normal frames in HR-Crime for testing is shown in the bottom two rows. None of the *Abuse* videos is annotated with ground truth labels by the authors of [16].

	Abuse	Arrest	Arson	Assault	Burglary	Explosion	Fighting	Road Accidents	Robbery	Shooting	Shoplifting	Stealing	Vandalism	**Total anomalous**	**Total normal**
Videos	50	50	50	50	100	50	50	150	150	50	50	100	50	**950**	**950**
HR-videos	38	42	48	47	96	26	39	68	145	46	50	98	46	**789**	**782**
Test HR-videos	0	5	9	3	13	12	5	13	5	21	21	5	5	**117**	**122**
Anomalous frames	0	7820	8166	8520	16104	3097	4437	1233	3815	9314	7525	6007	2078	**78116**	**0**
Normal frames	0	25804	19722	18041	60420	31525	7855	10922	4512	65165	68609	13804	8999	**335378**	**485227**

2.2 Feature Extraction

The feature extraction pipeline starts with extracting human proposals from each frame of the video and extract skeletons from each proposal. As shown in Fig. 1, first, human body proposals are detected using YOLOv3-spp, a Dense Connection and Spatial Pyramid Pooling Based version of YOLOv3 [5]. Out of human body proposals, skeletons are retrieved employing AlphaPose [3]. Each skeleton consists of 17 keypoints representing body part locations. Skeletons are later tracked using PoseFlow [17] which uses FLANN and ORB descriptors to match two consecutive frames. The model we employed then uses tracked information in the last default 100 frames and using Hungarian algorithm [10] solves the bipartite matching problem for the current frame. PoseFlow uses Non-

Maximum Suppression just as AlphaPose to remove redundant trajectories and rematch disjoint pose flows. Given an input video, the skeletons are tracked over the frames to model the trajectory of each person appearing in a given video. The trajectories are later converted to CSV-files, which can be inputted to MPED-RNN human-based anomaly detector.

3 Experiments

To evaluate MPED-RNN, authors in [12] mainly used the HR-ShanghaiTech dataset [12]. This dataset originally does not contain ground-truth labels for human bounding boxes, skeletons or trajectories. Also, the feature extraction step on HR-ShanghaiTech is missing from the MPED-RNN pipeline provided by the authors. Therefore, we were unable to evaluate the feature extraction steps separately on the HR-ShanghaiTech. Thus, we propose to compare the performance of the pre-trained MPED-RNN on the trajectories obtained using our feature extraction pipeline against the trajectories provided by the authors. For completeness of the proposed dataset, we also include the baseline results of the MPED-RNN network applied on the skeleton trajectories of the HR-Crime.

3.1 Datasets

The HR-ShanghaiTech [12] consists of 101 human-related videos with anomalies such as running, jumping, riding a bicycle, etc. against the normality which is simply walking. HR-ShanghaiTech is a subset of ShanghaiTech dataset [11] consisting of 107 videos filmed by 12 cameras on the Shanghai Tech University campus.

Both the HR-ShanghaiTech test set [12] and UCF-Crime [16] test set are provided with frame-level ground truth labels by the authors indicating if the event is anomalous or normal. Table 1 shows that 239 videos of the UCF-Crime testing set are human-related (the sum of 117 anomalous and 122 normal videos). Only these videos are used for the evaluation since the remaining videos are not labelled at frame-level.

3.2 Results and Discussion

Following [12], we use the frame-level Receiver Operating Characteristic (ROC) curve and its corresponding Area Under (AUROC) to evaluate the performance of the methods on the HR-ShanghaiTech and HR-Crime datasets. A higher AUROC value indicates better performance.

A) Pose extraction HR-Crime: Figure 4 shows a few examples of skeletons extracted from HR-Crime using our feature extraction pipeline discussed in Sect. 2. As can be seen, the skeletons are reasonably accurate for higher quality videos and where the person is clearly in the camera view. However, for videos of lower quality or insufficient lighting, such as in Fig. 4b, the feature extraction pipeline fails to detect people and their poses.

B) New baseline for HR-ShanghaiTech: Authors of [12] made available 12 models, each trained separately on a camera subdivision of HR-ShanghaiTech. HR-Crime however, is not structured in a camera wise manner. Therefore to ensure further consistencies, we trained the MPED-RNN architecture, *de novo*, on the whole HR-ShanghaiTech training set from all the 12 cameras. For this training, we still used the trajectories provided by the authors. The model obtained from training on the entire HR-ShanghaiTech training set achieves a slightly lower performance in AUROC: 0.735 (see Table 2) compared to the 0.754 reported in [12], where 12 models are trained on the 12 camera subdivisions of HR-ShanghaiTech. This indicates that camera settings influence the complexity of the anomaly detection problem, at least when using MPED-RNN.

Moreover, as the authors of MPED-RNN did not provide detailed information on feature extraction steps from the videos, it is not possible to determine if feature extraction was purely done automatically or if human input was involved in refining the extracted features. This might be an important reason why using our feature extraction pipeline performs less accurately than the original trajectories given in [12] (0.534 AUROC score as compared to 0.735 AUROC score obtained by [12]).

C) MPED-RNN on HR-Crime: Our first approach to establish the HR-Crime baseline is to test the pre-trained MPED-RNN with the entire HR-ShanghaiTech training set (on trajectories provided by [12]), on the HR-Crime test set, without explicitly fine-tuning the model to the new domain. We report the performance on HR-Crime class-wise as well as on the entire test set. This shows how well the information learned from the HR-ShanghaiTech can generalize to other categories of crimes scenes.

The obtained results are presented in Table 2 and Table 3. They are reported based on the frame level reconstruction and prediction anomaly scores obtained by the MPED-RNN models. We observe that the AUROC scores are highest for videos of the type *Assault* (0.7487) and *Stealing* (0.7337). In contrast, *Arson* and *RoadAccidents* have the lowest AUROC scores, 0.4290 and 0.4171 respectively. These results indicate that the pre-trained MPED-RNN model can make promising predictions even on an unseen domain if the 'human' subjects are fairly present in committing the anomalies. This is also an indication of suitability of the proposed feature extraction pipeline.

D) Fine-tuning MPED-RNN on HR-Crime: The second approach for establishing the HR-Crime baseline is training the MPED-RNN architecture to the newly created HR-Crime. For this purpose, we train the MPED-RNN model *de novo* on the HR-Crime dataset. For completeness, we also propose to fine-tune the pre-trained MPED-RNN model on the HR-ShanghaiTech dataset, on HR-Crime. With this approach, we aim at closing the structural gap between the two datasets. The results are presented in Table 2. The model trained *de novo* on HR-Crime achieves 0.6030 AUROC performance which is explained by the increased complexity of the HR-Crime dataset as compared to HR-ShanghaiTech. Surprisingly, the fine-tuned model achieves a lower AUROC performance of 0.5879 as compared to the model trained *de novo*. We suspect, the information gain from

(a) Satisfactory feature extraction

(b) Unsatisfactory feature extraction

Fig. 4. Visual examples of skeletons (in red) obtained from HR-Crime videos using our feature extraction pipeline. The first row displays examples of satisfactory extractions from clear videos. The bottom row shows examples of unsatisfactory extractions.

Table 2. Baseline results for the HR-ShanghaiTech and HR-Crime datasets using MPED-RNN models.

Dataset	AUROC
B) New baseline for HR-ShanghaiTech	
HR-ShanghaiTech	**0.7346**
C) MPED-RNN on HR-Crime	
HR-ShanghaiTech (pre-trained) [12]	0.5346
D) Fine-tuning MPED-RNN on HR-Crime	
HR-Crime (de novo)	**0.6030**
HR-ShanghaiTech (fine-tuned)	0.5879

Table 3. Class-wise results for the HR-Crime dataset using the pre-trained MPED-RNN on HR-ShanghaiTech.

Dataset	Crime class	AUROC
C) MPED-RNN on HR-Crime		
HR-ShanghaiTech (pre-trained) [12]	Arrest	0.5617
	Arson	0.4290
	Assault	0.7487
	Burglary	0.6790
	Explosion	0.4740
	Fighting	0.4847
	Road accidents	0.4171
	Robbery	0.6586
	Shooting	0.4900
	Shoplifting	0.6342
	Stealing	0.7337
	Vandalism	0.6396

the HR-ShanghaiTech model does not generalize well to the HR-Crime dataset. This emphasizes the complexity gap between the two datasets that is apparent when comparing anomalies such as running and jumping in HR-ShanghaiTech to the anomalous events in HR-Crime. Comparing to the results of the experiment (C), our fine-tuned model shows a slight increase in performance which might also suggest that the pre-training approach is relevant to consider when comparing two datasets with distinct structures.

4 Conclusion

In this work, we discussed the preparation steps to develop HR-Crime, and provided a baseline human-related anomaly detection analysis on it. The dataset as well as the feature extraction pipeline will be publicly available for further use by the community.

The results presented in Sect. 3 (C) show that the pre-trained MPED-RNN on HR-ShanghaiTech does not perform as well on HR-Crime compared to the performance achieved on HR-ShanghaitTech itself. We suspect this to happen due to the complexity gap between the two datasets as also has been observed in other domains [13] and [18]. HR-ShanghaiTech consists of videos shot on the same University campus. HR-Crime, on the other hand, is a collection of YouTube videos, where each video is filmed in a different location. Also, the types of anomalies differ greatly: HR-ShanghaiTech contains anomalous events such as *running* and *jumping*, while HR-Crime consists of real-world crime scenes with natural movements that are not staged. The quality of the videos also plays an important role in anomaly detection. The HR-ShanghaiTech videos are of high quality, while the HR-Crime videos range in quality and lighting. Thus, skeletons and trajectories are detected less accurately for HR-Crime.

Another factor that plays an essential role in the high number of false negatives for anomalous events is the frame-level evaluation. Each test video in HR-Crime is annotated with ground truth labels indicating the window of an anomalous event. These windows indicate the start and end frame of the event. As mentioned before, the HR-Crime dataset is composed of a complex set of videos. Therefore an anomalous event such as a *Burglary* can have multiple anomalous and normal movements in the same video. However, prediction and evaluation are made per frame. An alternative that we plan on exploring is to label the videos temporally and spatially to not only evaluate if a frame contains an anomalous event but also to find the area where it occurs as suggested in [14]. This is a more accurate way of evaluating anomalous events. However, it requires laborious work to annotate video frames manually. Future lines of research will also be dedicated to the categorical classification of the identified anomalies through the analysis of the descriptors of the movement of the human body.

References

1. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. PAMI (2019)
2. Emonet, R., Varadarajan, J., Odobez, J.M.: Multi-camera open space human activity discovery for anomaly detection. In: IEEE International Conference on AVSS (2011)
3. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: RMPE: regional multi-person pose estimation. In: IEEE International Conference on Computer Vision (2017)
4. Gong, M., Zeng, H., Xie, Y., Li, H., Tang, Z.: Local distinguishability aggrandizing network for human anomaly detection. Neural Netw. **122**, 364–373 (2020)

5. Huang, Z., Wang, J., Fu, X., Yu, T., Guo, Y., Wang, R.: DC-SPP-YOLO: dense connection and spatial pyramid pooling based yolo for object detection. Inf. Sci. (2020)
6. Insafutdinov, E., et al.: ArtTrack: articulated multi-person tracking in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
7. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 34–50. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_3
8. Iqbal, U., Milan, A., Gall, J.: PoseTrack: joint multi-person pose estimation and tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
9. Kocabas, M., Karagoz, S., Akbas, E.: MultiPoseNet: fast multi-person pose estimation using pose residual network. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 437–453. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01252-6_26
10. Kuhn, H.W.: The Hungarian method for the assignment problem. Naval Res. Logistics Q., 83–97 (1955)
11. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection-a new baseline. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
12. Morais, R., Le, V., Tran, T., Saha, B., Mansour, M., Venkatesh, S.: Learning regularity in skeleton trajectories for anomaly detection in videos. In: IEEE Conference on Computer Vision and Pattern Recognition (2019)
13. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 413–420. IEEE (2009)
14. Ramachandra, B., Jones, M.: Street scene: a new dataset and evaluation protocol for video anomaly detection. In: IEEE Winter Conference on Applications of Computer Vision (2020)
15. Ramachandra, B., Jones, M., Vatsavai, R.R.: A survey of single-scene video anomaly detection. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
16. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
17. Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C.: Pose flow: efficient online pose tracking. arXiv preprint arXiv:1802.00977 (2018)
18. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1452–1464 (2017)