The Concept of the Deviant Behavior Detection System via Surveillance Cameras

Nikolay Teslya[1,2], Igor Ryabchikov[2(✉)], and Evgeniy Lipkin[2]

[1] SPIIRAS, 14th Line 39, 199178 St. Petersburg, Russia
teslya@iias.spb.su
[2] ITMO University, Kronverksky pr. 49, 197101 St. Petersburg, Russia
i.a.ryabchikov@gmail.com, eugenejsm@mail.ru

Abstract. One of the tasks of a Smart City is an ensuring the safety of its residents. Dangerous situations are often caused by deviant behavior of people: robbery, fight, kidnapping, attack, etc. Nowadays all cities are equipped with video surveillance systems that are used to monitor city life. They also can be used to detect potentially dangerous situation in a real-time and to inform emergency services to taking actions that will help to eliminate them. Due to the huge amount of video material, to solve the problem of detecting dangerous situations, it is necessary to use modern intelligent technologies that allow their automatic analysis. This paper proposes a review of modern technologies, and concept of the deviant behavior detection system based on the reviewed technologies.

Keywords: Deviant behavior · Surveillance · Computer vision · Smart city · Neural network · 3D skeleton · Markov logic networks

1 Introduction

Nowadays many cities are increasingly introducing the concept of a smart city to ensure sustainable development. Within the scope of this concept, the use of various sensors and data accumulation and processing systems for the analysis and formation of decisions and recommendations on the management of the city is considered. Also, gathered information in a smart city can be used to provide services for the city citizens, making life in the city more comfortable.

One of the important part of smart city system is a Closed-circuit television (CCTV) or video surveillance system. For example there are more than 500000 CCTV cameras only in London. The CCTV is mostly used now for ensuring the safety of smart city residents. Until recently, it was used mainly as evidences sources and for analyzing incidents after their completion. But it has much greater potential, which can be revealed through the use of modern computer vision technologies, which allow automatic analysis of a huge amount of video in real time. An example of the

This work was partly supported by the grant of ITMO University (Project № 618268).

© Springer Nature Switzerland AG 2019
S. Misra et al. (Eds.): ICCSA 2019, LNCS 11624, pp. 169–183, 2019.
https://doi.org/10.1007/978-3-030-24311-1_12

successful application of such technologies is the face recognition system which was used at the 2018 World Cup and helped to identify over 180 wanted offenders [1].

But computer vision technologies are not limited to face recognition – they can provide the detection of complex scenes, thereby providing real-time awareness of the dangerous situations occurring in the city and allowing timely actions to eliminate them. Often a dangerous situation is caused by the deviant behavior of people: robbery, assault, fighting, placing suspicious objects in public places, etc.

The complexity of the task of deviant behavior detection lies in the strong variability of scenes of interest: they can run for a variable amount of time and contain a variable number of people performing a wide range of interactions. At the same time, the amount of data that can be utilized by machine learning techniques (examples of scenes of interest) is relatively small, taking into account the variability. For example, in accordance with recent burglary published in UK Police Department web site [2], after getting money from ATM, a victim was followed by two individuals, and then stopped with threatening and demanding cash. The victim managed to run off but the men caught up with him and subjected him to a prolonged violent attack during which he was punched and kicked repeatedly and threatened with knives. His attackers then ran off.

The paper presents an overview of the existing intelligent technologies that can be used to develop a system for detecting complex scenes of deviant behavior via the city's video surveillance cameras, and the concept of such a system. Section 2 presents adopted systems of dangerous situations detection in smart cities. Section 3 describes the early approaches of solving the problem of deviant behavior detection. Section 4 presents an overview of modern computer vision technologies, which make it possible to obtain a compact set of sufficiently representative features of the scene. Section 5 presents examples of scenes of deviant behavior and their distinguishing features that can be detected. Section 6 describes the proposed concept of a deviant behavior detection system.

2 Dangerous Situations Detection in Smart Cities

Timely detection of dangerous situations will make it possible to provide immediate help to victims, prevent further escalation and detain the perpetrators. The use of sensors and application of intelligent technologies for this purpose falls into the concept of smart city. At the moment there are a number of successful projects aimed at detecting dangerous situations:

- Detection and localization of gunfire using sound sensors. The technology allows not only to respond quickly to the incident, but also to provide the police with a vision of the situation and develop a strategy for action. An example of its implementation is the ShotSpotter project [3], adopted in more than 90 cities of USA.
- Wanted criminals detection with face recognition technologies. An example of the implementation is the FindFace Public Safety system [4], which was used at the 2018 World Cup and helped to identify over 180 wanted offenders.

- Smoke detection via surveillance cameras. An example is the SmokeCatcher system [5], mainly used on industrial sites.
- Detection of abandoned objects in public places via surveillance cameras. An example is the Reco3.26 system [6].

But so far no finished system has been introduced for detecting deviant behavior of people that is ready for use in a real environment. A review of existing approaches is presented below.

3 Early Solutions for Deviant Behavior Detection

In earlier works, the task of classifying people's activities was mainly divided into three stages: segmentation of regions of interest in frames based on simple hypotheses (for example, by the presence of motion); handcrafted features calculation for the selected regions; and video clips classification according to these features using machine learning, for example, the Support Vector Machines (SVM) classifier. For instance, in [7], an approach for people's activities classification based on optical flow analysis, histograms of oriented gradients and the Bag-of-Words (BoF) model utilization is presented. First, the points of interest are detected by the scale-invariant features transformation (SIFT) algorithm. Then, histograms of oriented gradients and histograms of optical flow, which are used as words in the Bag-of-Words model, are calculated in the surrounding regions of the points. Finally, SVM is used for the classification of activities in the frame.

The disadvantage of this approach in relation to the task of deviant behavior detection is the focus on detecting only activities that have distinctive signs of movement (speed, acceleration). But in complex scenes of deviant behavior, there may be no obvious differences in speed and acceleration (for example, a robbery without a fight or a confrontation scene preceding a fight). In addition, the handcrafted features often have poor interpretability and reflect actions of people only indirectly, due to which the behavior of such classifiers can be unpredictable when used in a real diverse environment. But despite the fact that at the moment there are more advanced computer vision technologies that allow for a more in-depth analysis of videos, such approaches are still emerging, improving the results of earlier work. The main motivation is the fact that these solutions are often less computationally expensive.

In [8], a fight detection approach is proposed that is largely similar to [7] but differs by the method of region segmentation - segmentation is based on the optical flow instead of SIFT algorithm, also there are differences in histograms of oriented gradients and optical flow calculation. Similarly, the Bag-of-Words model is applied to histograms and the classification is performed using SVM. The approach was tested on a dataset containing scenes of hockey fights [9] and archived 95% accuracy.

In [10], an approach of violent behavior detection in crowds based on the analysis of changes in textural properties of the Gray Level Co-occurrence Matrix (GLCM) over time is proposed. First, each frame of the video is divided by the grid into multiple sub-frames of equal size. Further processing and classification are performed individually for each sub-frame. To remove the effect of the background, the difference of the

GLCM matrices for adjacent frames is calculated. Then, based on the resulting matrix, various statistical properties are calculated: energy, contrast, homogeneity, correlation, and dissimilarity. According to the properties obtained for each time frame of the video, the final properties are calculated: mean, standard deviation, skewness, inter-frame uniformity. Then the classification is performed using Random Forests. The method was tested on the Violent Flow dataset [11] and showed 86% accuracy (for comparison, MoSIFT [7] showed 89% accuracy).

With the development of convolutional neural networks, a host of computer vision technologies have emerged that allows the estimation of various features of a video content giving a glance on what is happening. But so far no solution has been proposed that uses their full potential. An example is the Eye in the Sky project, which proposes the use of drones for observing areas and detecting fights based on a 2D skeleton (key points of the body) of a person [12]. For fights classification, SVM is used, which consumes angles between the limbs of a body of a single person. The proposed approach achieved 88.8% accuracy on the data set proposed by the developers. The main shortcomings of the approach is that the system takes into account only a single person pose, and does not consider a movement in time.

4 Related Modern Computer Vision Technologies

The main difficulty in detecting scenes of deviant behavior of people lies in the strong variability of such scenes. To reduce variability, it is necessary to form a small but representative set of properties describing the observing scene for further analyses. Below are the technologies that allows the extraction of such properties from plain RGB videos.

4.1 Detection of People

Detection of people is one of the first tasks that must be solved for the purpose of deviant behavior classification. Detection refers to the evaluation of bounding boxes for each person present in the frame.

The best results for the detection of pedestrians within the MOT17Det Challenge (https://motchallenge.net/data/MOT17Det/) were achieved in works [13–15]. All of them offer variations of convolutional neural networks with modifications to solve the multi-scale problem.

4.2 Tracking and Re-identifying People on Video

Another vital task for people's actions analyzation is matching of people detected on individual frames. To solve it, various properties of appearance and behavior (colors, speed, etc.) are used, as well as the assumption of smooth movements of people in adjacent frames, manifested in a slow change in their position and posture. The difficulty this task acquires in crowded scenes with a significant overlap of people. The popular challenges for tracking people on video are Multi-Person PoseTrack 2018

(https://posetrack.net/workshops/ecc201201/) and VOT2018 (http://www.votchallenge.net/vot2018/). In Multi-Person PoseTrack, in addition to short-term tracking of people, there is the task of two-dimensional poses evaluation. In VOT2018, the task is to assess the bounding boxes of people during the short-term and long-term tracking (when a person can disappear from view for a long period of time). The best results on Multi-Person PoseTrack were obtained in [16], which proposes the propagation of poses from previous frames using optical flow, and their comparison with the detected ones. In [17], the results of the VOT2018 Challenge and description of the participants' approaches are presented.

4.3 Three-Dimensional Human Skeleton Evaluation

Probably the most representative properties of the scene are three-dimensional human skeletons (Fig. 1) - sets of key points of bodies (head, torso, shoulders, elbows, palms, pelvis, knees, feet, etc.). On their basis, a large range of actions and interactions of people can be classified, for example, punches, kicks, transfer of an object, search of pockets, etc.

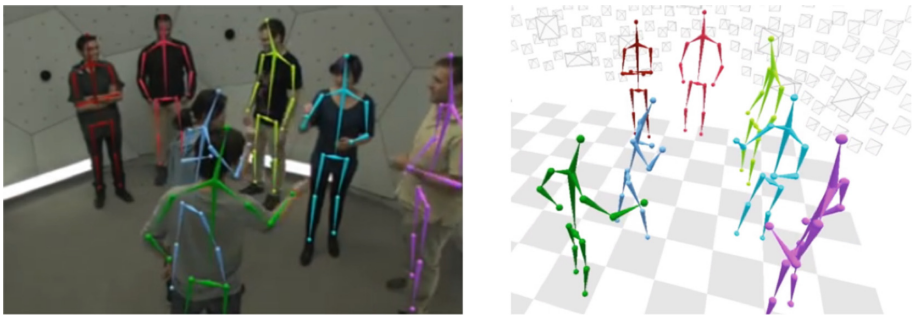

Fig. 1. Example of three-dimensional skeletons evaluation from the CMU Panoptic dataset [16].

One of the most popular data sets for training and validating human actions detection models is Human3.6M [19], associated with the prestigious challenge ECCV 18 (https://posetrack.net/workshops/eccv2018/). This challenge proposes a subset of images of Human3.6M and sets the task of detecting the three coordinates of each joint in the 3D space of the camera. The best results were achieved in [20–23]. A survey of human actions classification approaches using skeletons is presented in [24].

4.4 Objects Segmentation and Classification

A scene of deviant behavior is characterized by the interaction of people with each other and surrounding objects. Such objects can be bags, shop windows, ATMs, vehicles, tools, etc. Thus, a classifier of objects is an important component for understanding the scene. At the same time, the segmentation of objects is necessary to determine the exact relative position of objects and people, as well as the dimensions of objects. In this case, segmentation refers to the evaluation of areas directly occupied by

objects. The task of segmentation and classification rises in the popular challenge COCO Detection 2018 (http://cocodataset.org/#detection-2018). Best segmentation results were achieved using Mask R-CNN [25]. A review of modern approaches for the classification of objects is presented in [26].

4.5 Surfaces Segmentation and Determination of Their Orientation

The landscape determination (surrounding obstacles, walls, roads, etc.) can help in analyzing the actions of people on a scene. For example, determining the location of the walls and correlating it with the location of the alleged victim and deviants, it can be concluded whether the victim could retreat or was blocked, which in turn increases the likelihood of a deviant behavior.

In [27, 28], approaches are proposed for segmentation of surfaces (planes) and determination of their orientation using a single RGB image. The approaches are based on the use of neural networks, trained end-to-end. For training, data sets containing RGB-D images are used, allowing to segment and estimate the parameters of the presented three-dimensional planes with simple calculations. Neural networks take an RGB image as input and provide segmentation maps (probabilities of each pixel belonging to one or another plane) and parameters of these planes (three-dimensional normal vectors) as output.

4.6 Hand Gestures Detection

Specific gestures are also characteristic of various participants (victims, aggressors) in a deviant behavior scene. Simple examples are clenched hands, grabbing a person by clothes, or, conversely, trying to prevent an attack with an open palm. A review of papers on the classification of hand gestures is presented in article [29].

4.7 Face Detection

Currently, there are numerous technologies for face detection and individuals identification, for example, the popular open source system OpenFace [30]. In video surveillance, an example of the use of such technologies is the identification of wanted criminals. For the purpose of deviant behavior detection, it is proposed to use these technologies to determine whether the face of the suspect is open (and it is possible to use the image for person identification in future). If the face of the suspect is covered with a mask or helmet, this increases suspiciousness and may testify in favor of the deviant behavior manifestation.

4.8 Gender Recognition

Gender can increase confidence in the classification of ambiguous cases, given the statistics of manifestations of deviant behavior (for example, the likelihood of a woman being a robber may be less than a man [31]). In paper [32], the Faster-RCNN convolutional neural network for gender classification, pre-trained for human detection, is

proposed. The work [33] proposes the use of SVM for gender classification by features of color and texture—Hue-Saturation-Value (HSV) and Local Binary Patterns (LBP).

4.9 Determination of Clothing

Clothing can also be used as an additional feature when classifying ambiguous scenes, reflecting the "aggressiveness" of the scene participants. For that purpose, the color range and type of clothing may be considered. The work [34] proposes the calculation of color (HSV) and textural (HOG, MB-LBP) features of clothing items, and the subsequent classification of the type of clothing using Gentle AdaBoost [35] and color using KNN. The work [36] also proposes the calculation of various color (histograms of colors) and textural features (HOG, Bag of dense SIFT Features, DCT responses). SVM is used for clothing classification. In [37] the use of a convolutional neural network (based on AlexNet) for classifying the type of clothing and defining various attributes (color, sleeve length, neckline, etc.) is proposed.

5 Examples of Deviant Behavior Scenes

Using the computer vision technologies described above, a lot of features of observing scenes can be extracted: the position and appearance of people, the surrounding objects, the landscape, the interaction of people with each other and with objects. We assume that these features are sufficient for detecting a large spectrum of deviant behavior scenes. For example, consider the scenes presented in the UCF-Crime [38] and I-LIDS [39] data sets. UCF-Crime contains 1900 videos (128 h in total), which depict 13 types of scenes of abnormal incidents and deviant behavior of people recorded by surveillance cameras: abuse, arrest, arson, assault, road accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting, vandalism. Videos were collected from YouTube and LiveLeaks. I-LIDS is provided by the UK government to benchmark Video Analytics systems. It focuses on several tasks, in particular, the detection of abandoned packages, the observation of doorways (detection of all incoming/outgoing people) and tracking of people through several intersecting and non-intersecting video surveillance cameras. For each scenario, there is approximately 24 h of video.

Most of the fight scenes from the UCF-Crime set correlate with the presence of simple intrinsic actions (in particular, punches and kicks) performed towards a near person, falls and specific poses (Fig. 2 left and middle). These actions are short in time, highly discriminative, and can be distinguished observing a three-dimensional human skeleton. But of particular interest is the task of fights detection before their escalation, when timely intervention can prevent them. In that case, individual actions performed by people are less expressive, but when considering actions of all the participants in the scene for a sufficiently long period of time, it is possible to distinguish suspicious activity. For example, in Fig. 2, the right represents a frame of a confrontation scene without a clear manifestation of violence. Deviant behavior is manifested in the following actions: the advance and retreat of the parties of the conflict, the pursuit while keeping each other in sight, accompanied by grasping and specific gestures. For the

detection of such scenes, the following technologies can be used: tracking of people, three-dimensional skeletons evaluation and hand gestures classification.



Fig. 2. Fight scenes from the UCF-Crime dataset.

The important task is the detection of abandoned packages left in public places (Fig. 3). For that purpose, we need to track people and objects carried by them (in the hands, on the back, etc.) and detect when the carrier moves away from the object (while taking into account special cases, such as transfer of the object to another carrier). We can also detect the sudden absence of the object in a person (if the fact of placing the object did not get on the video) and the presence of a placed object, near which there is no permanent person. For that task, we can use technologies for objects segmentation and classification, tracking and re-identification of people and objects, three-dimensional skeletons evaluation.



Fig. 3. The scene of leaving the bag in a public place from the I-LIDS dataset.

Another type of deviant behavior is a burglary of ATMs (as well as cars, shop windows, etc.) (Fig. 4). Its characterizing signs can be the presence of people near ATMs in non-standard positions (bent, in a squat, etc.) or performing some “active” movements in relation to the ATM (strikes, using a lever, etc.). Increases the likelihood of deviant behavior the presence of objects in hands of suspects. To avoid false positive

detections during regular maintenance, we can classify the appearance (form) of the employee, as well as check the scheduled service time. Technologies for people tracking, three-dimensional skeleton evaluation, segmentation and classification of objects (in the hands of people, ATMs, etc.) and clothing classification can be used.



Fig. 4. The ATM robbery scene from the UCF-Crime dataset.

Many types of robbery can be recognized by detecting the pose of the threat of a weapon against a nearby person (Fig. 5), in which a robber is in for a sufficiently long period of time. To detect the pose, technologies for three-dimensional skeletons evaluation, objects segmentation and classification (in the hands of an intruder) can be used. Increase confidence in the correct classification of the threat can the hidden face of the attacker (face detection technology).



Fig. 5. Scenes of robbery with the threat of a weapon from the UCF-Crime data set.

A popular method of robbery is that one or several people drive up to a victim, search/retrieve items (bags, wallets, etc.) and quickly drive away from the crime scene (Fig. 6). Sometimes a robbery is accompanied by the threat of a weapon. To detect this type of robbery, we can track the approaching of strangers, the transfer of objects to

them (or the search of pockets) and their escape. As well as additional characteristic actions of alleged victims and robbers (did the victim try to retreat, were there any signs of resistance). To detect such scenes we can utilize technologies for tracking of people and objects, three-dimensional skeleton evaluation, segmentation and classification of objects, as well as technologies for detecting surfaces and their orientations (for correlating the actions of participants of the scene with the terrain, for example, whether the victim could retreat or was behind the wall).



Fig. 6. Scenes of the robbery by people on vehicles from the UCF-Crime dataset.

6 The Concept of the Deviant Behavior Detection System

For deviant behavior scenes classification it is tempting to use machine learning to automatically build a model. But, although the computer vision technologies described above can significantly reduce the variability of the original set of features (RGB frame sequences), the scenes of interest still have strong variability, since they can last a variable amount of time and contain a variable number of interacting people and objects. In addition, it's difficult to get a sufficient number of realistic examples of deviant behavior sequences to train the model. The solution is the utilization of crime experts knowledge. Based on the gathered knowledge, descriptions of scenes of interest will be formed in a formal language that supports inference (for instance, first-order logic). The elementary units (atoms) will be individual simple actions of people in time (sit down, take an object, get closer to a person or an object, etc.), features of people's appearance (gender, clothes), features of objects (type, size) and features of the terrain (wall positions, time of a day, terrain type) calculated on the basis of the computer vision components outputs and the provided background knowledge. With this approach, the decisions taken by the detection system can be explained. In [40], the application of first-order logic for scene recognition by smart agents in smart homes is considered. In the proposed approach, first, various qualitative features of objects in the room are detected (shape, color, topology and location) as well as the presence of special target objects that the agent can classify (TV, printer, table, etc.). Features are detected by low-level perception components and are represented as predicates used during reasoning in accordance with the agent's domain knowledge. As a result of the reasoning, the observed picture can be supplemented, and conclusions can be made about the condition of the room and the necessity of intervention (for example, spilled water in a room near an open window when it rains may indicate the need to shut the window and wipe the floor).

As an example, let's consider a possible description of the scene of the robbery presented in Fig. 6 in a first order logic (Table 1). For simplicity, this description of the scene does not take into account the time or sequence of events, but it's taken into account in atoms formed on the basis of the outputs of the perception components at a lower level (in particular, DistanceReducedBetween (Drives, Vict) and DistanceIncreasedBetween (Drives, Vict)). It should be noted that confidence in detecting deviant behavior can be increased by considering additional signs, in addition to the rules presented in Table 1. Among such signs are weapons threatening, victim's attempt to retreat, signs of struggle, gender of participants, the presence of masks on intruders, etc. Moreover, almost all basic features can be detected only with a certain confidence due to the specificity of the machine perception components and the fuzziness of certain features. Thus, the use of a probabilistic model suggests itself.

Table 1. Description of a street robbery scene in a first order logic.

№	Knowledge base rules
1	∀Vehic.∀Vict. ∀Crim. DistanceReducedBetween(Vehic, Vict) ∧ (GotOutOfA(Crim,Vehic) ∨ WasOnA(Crim,Vehic)) ∧ WereClose(Crim, Vict) → DroveUpTo(Crim, Vict)
2	∀Crim. ∀Vict. ∀Bag. ReachedForAPocket(Crim, Vict) ∨ ReachedForABag(Crim, Bag) ∧ HoldABag(Vict, Bag) → SearchedA(Crim, Vict)
3	∀Vict. ∀Item. ∀Crim. HoldedA(Vict, Item) ∧ HoldedA(Vict, Item) ∨ WereInHandingPose(Vict) → TookAnItem(Crim, Vict)
4	∀Vehic. ∀Vict. ∀Crim. DistanceIncreasedBetween(Vehic, Vict) ∧ (GotInA(Crim,Vehic) v WasOnA(Crim,Vehic)) ∧ WereClose(Crim, Vict) → DoveAwayFrom(Crim, Vict)
5	∀Vehic.∀Vict. ∀Crim. DroveUpTo(Crim, Vict) ∧ (SearchedA(Crim, Vehic) ∨ TookAnItem(Crim, Vict)) ∧ DroveAwayFrom(Crim, Vict) → RoberryHappend(Crime, Vict)

The integration of probabilistic models and relational structures is the main topic of Statistical Relational Learning (SRL) discipline. It provides the formation of complex knowledge structures, while taking into account the various uncertainties of the real world. One of the most popular representational formalisms of SRL is the Markov Logic Network (MLN) [41], combining first-order logic and probabilistic graphical models (Markov networks). MLN is a first-order knowledge base with weights associated with each formula and reflecting their veracity. Through the grounding of the formulas of the knowledge base, a Markov network is formed that contains nodes for each ground atom and edges between the nodes if the corresponding ground atoms are found together in at least one ground formula. Inference is carried out on the formed Markov network. The disadvantage of MLN is the computational complexity of training and inference in Markov Networks, but there are approaches that allow you to get closer to the solution with reasonable computational costs. An example of the use of MLN in computer vision is presented in [42], which poses the task of matching people and their cars in parking lots via surveillance cameras.

The concept of the deviant behavior detection system proposed in this paper is presented in Fig. 7. It is proposed to form an initial set of features of the observed scene

based on the technologies for people detection and tracking, three-dimensional skeleton evaluation, segmentation and classification of objects, segmentation and orientation evaluation of surfaces, gestures classification, face detection, gender and clothing classification. Further, the high-level scene properties are formed in the form of predicates of first-order logic. To classify the actions and postures of people, it is proposed to use approaches based on machine learning, for example, [43], most of the other properties can be hand-crafted. For example, to determine the predicate DistanceReducedBetween (Vehic, Vict), we can track the change in the distance between each pair of people and vehicles in the scene. For the classification of deviant behavior, scenes of interest will be modelled in a first order logic language based on the crime experts knowledge. When classifying complex scenes (containing many interrelated factors), MLN is suggested. The weights of the formulas will also be set by experts, and the weights of atomic predicates will be formed based on the detection confidence.

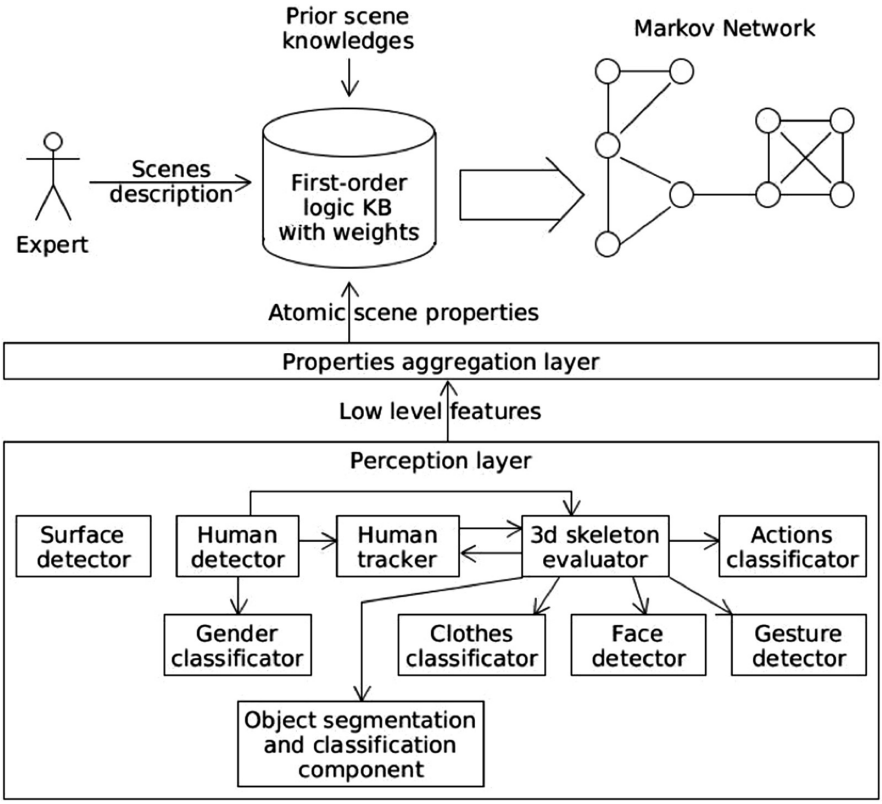

Fig. 7. Diagram of the deviant behavior detection system concept.

7 Conclusion

In this paper, the concept of a deviant behavior detection system via the city's video surveillance cameras was presented. Such a system can increase the awareness of the city administration about the events taking place in it, promptly react to problem situations and, as a result, increase the safety of residents. A brief overview of modern computer vision technologies providing a set of representative features of the observed scene to be formed was presented. To solve the problem of the strong variability of the scenes of interest, it is proposed to utilize the experts' knowledge representing the description of scenes in a first-order logic. To account for real life uncertainties, MLN utilization is proposed. Also, examples of real deviant behavior scenes were presented, and a possible formal description of the street robbery scene was proposed.

References

1. Rostec: NtechLab Face Recognition System Prevents Sponsor's Trophy Theft at the 2018 FIFA World Cup. https://rostec.ru/en/news/ntechlab-face-recognition-system-prevents-sponsor-s-trophy-theft-at-the-2018-fifa-world-cup-/. Accessed 10 Mar 2019
2. West Yorkshire Police. Appeal Over Violent Street Robbery, Armley, Leeds. https://www.westyorkshire.police.uk/news-appeals/appeal-over-violent-street-robbery-armley-leeds. Accessed 10 Mar 2019
3. ShotSpotter. https://www.shotspotter.com/. Accessed 5 May 2019
4. FindFace Public Safety. https://findface.pro/ru/face-recognition-public-safety.html. Accessed 5 May 2019
5. SmokeCatcher, Video Smoke Detection for critical environments. http://www.araani.com/en/smokecatcher/. Accessed 5 May 2019
6. Smart Recognition System Reco3.26. https://www.reco326.com/index.php/en/. Accessed 5 May 2019
7. Chen, M., Hauptmann, A.: Mosift: Recognizing human actions in surveillance videos. Pittsburgh (2009). http://www.cs.cmu.edu/~mychen/publication/ChenMoSIFTCMU09.pdf. Accessed 10 Mar 2019
8. Zhou, P., Ding, Q., Luo, H., Hou, X.: Violence detection in surveillance video using low-level features. PLoS One **13**(10) (2018). https://doi.org/10.1371/journal.pone.0203668
9. Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R.: Violence detection in video using computer vision techniques. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds.) CAIP 2011. LNCS, vol. 6855, pp. 332–339. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23678-5_39
10. Lloyd, K., Rosin, P.L., Marshall, D., Moore, S.C.: Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures. Mach. Vis. Appl. **28**(3–4), 361–371 (2017). https://doi.org/10.1007/s00138-017-0830-x
11. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: real-time detection of violent crowd behavior. In: Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–6. IEEE (2012). https://doi.org/10.1109/cvprw.2012.6239348
12. Singh, A., Patil, D., Omkar, S.N.: Eye in the sky: real-time drone surveillance system (DSS) for violent individuals identification using scatternet hybrid deep learning network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1629–1637 (2018)

13. Cai, Z., Fan, Q., Feris, Rogerio S., Vasconcelos, N.: A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 354–370. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_22
14. Yu, F., Li, W., Li, Q., Liu, Yu., Shi, X., Yan, J.: POI: multiple object tracking with high performance detection and appearance feature. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 36–42. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_3
15. Lin, C., Lu, J., Wang, G., Zhou, J.: Graininess-aware deep feature learning for pedestrian detection. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 745–761. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_45
16. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 472–487. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_29
17. Kristan, M., et al.: The sixth visual object tracking VOT2018 challenge results. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11129, pp. 3–53. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11009-3_1
18. Joo, H., et al.: Panoptic studio: a massively multiview system for social motion capture. In: IEEE International Conference on Computer Vision, pp. 3334–3342. IEEE (2015). https://doi.org/10.1109/iccv.2015.381
19. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6 m: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014). https://doi.org/10.1109/tpami.2013.248
20. Sárándi, I., Linder, T., Arras, K.O., Leibe, B.: Synthetic occlusion augmentation with volumetric heatmaps for the 2018 ECCV posetrack challenge on 3D human pose estimation (2018). https://arxiv.org/pdf/1809.04987.pdf. Accessed 10 Mar 2019
21. Sun, X., Li, C., Lin, S.: An Integral Pose Regression System for the ECCV 2018 PoseTrack Challenge. (2018). https://arxiv.org/pdf/1805.04095.pdf. Accessed 10 Mar 2019
22. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3D human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7307–7316. IEEE (2018). https://doi.org/10.1109/cvpr.2018.00763
23. Park, S., Kwak, N.: 3D Human Pose Estimation with Relational Networks (2018). http://bmvc2018.org/contents/papers/0530.pdf. Accessed 10 Mar 2019
24. Wang, P., Li, W., Ogunbona, P., Wan, J., Escalera, S.: RGB-D-based human motion recognition with deep learning: A survey. Comput. Vis. Image Underst. **171**, 118–139 (2018). https://doi.org/10.1016/j.cviu.2018.04.007
25. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: IEEE International Conference on Computer Vision, pp. 2961–2969). IEEE (2017). https://doi.org/10.1109/iccv.2017.322
26. Liu, L., et al.: Deep learning for generic object detection: A survey. (2018). https://arxiv.org/abs/1809.02165. Accessed 10 Mar 2019
27. Yang, F., Zhou, Z.: Recovering 3D planes from a single image via convolutional neural networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 87–103. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_6
28. Liu, C., Yang, J., Ceylan, D., Yumer, E., Furukawa, Y.: PlaneNet: piece-wise planar reconstruction from a single RGB image. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2579–2588. IEEE (2018). https://doi.org/10.1109/cvpr.2018.00273

29. Nandwana, B., Tazi, S., Trivedi, S., Kumar, D., Vipparthi, S.K.: A survey paper on hand gesture recognition. In: 7th International Conference on Communication Systems and Network Technologies (CSNT), pp. 147–152. IEEE (2017). https://doi.org/10.1109/csnt.2017.8418527
30. Baltrušaitis, T., Robinson, P., Morency, L.P.: Openface: an open source facial behavior analysis toolkit. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE (2016). https://doi.org/10.1109/wacv.2016.7477553
31. Rennison, C.M., Melde, C.: Gender and robbery: a national test. Deviant Behav. **35**(4), 275–296 (2014). https://doi.org/10.1080/01639625.2013.848104
32. Chahyati, D., Fanany, M.I., Arymurthy, A.M.: Man woman detection in surveillance images. In: 5th International Conference on Information and Communication Technology (ICoIC7), pp. 1–4. IEEE (2017). https://doi.org/10.1109/icoict.2017.8074682
33. Geelen, C.D., Wijnhoven, R.G., Dubbelman, G.: Gender classification in low-resolution surveillance video: in-depth comparison of random forests and SVMs. In: Video Surveillance and Transportation Imaging Applications 2015, vol. 9407, p. 94070 M. International Society for Optics and Photonics. SPIE, San Francisco (2015). https://doi.org/10.1117/12.2077079
34. Zhu, J., Liao, S., Lei, Z., Yi, D., Li, S.: Pedestrian attribute classification in surveillance: database and evaluation. In: IEEE International Conference on Computer Vision Workshops, pp. 331–338. IEEE (2013). https://doi.org/10.1109/iccvw.2013.51
35. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). In: The Annals of Statistics, vol. 28, № 2, pp. 337–407. Stanford (2000). http://statweb.stanford.edu/∼jhf/ftp/boost.ps. Accessed 10 Mar 2019
36. Yang, M., Yu, K.: Real-time clothing recognition in surveillance videos. In: 18th IEEE International Conference on Image Processing, pp. 2937–2940. IEEE (2011). https://doi.org/10.1109/icip.2011.6116276
37. Lao, B., Jagadeesh, K.: Convolutional neural networks for fashion classification and object detection. In: Computer Vision, CCCV 2015, pp. 120–129 (2016). http://cs231n.stanford.edu/reports/2015/pdfs/BLAO_KJAG_CS231N_FinalPaperFashionClassification.pdf. Accessed 10 Mar 2019
38. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6479–6488. IEEE (2018). https://doi.org/10.1109/cvpr.2018.00678
39. Branch, H.O.S.D.: Imagery library for intelligent detection systems (i-LIDS). In: IET Conference on Crime and Security, pp. 445–448. IET (2006). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/143870/ilids-brochure.pdf. Accessed 10 Mar 2019
40. Falomir, Z., Olteţeanu, A.M.: Logics based on qualitative descriptors for scene understanding. Neurocomputing **161**, 3–16 (2015). https://doi.org/10.1016/j.neucom.2015.01.074
41. Richardson, M., Domingos, P.: Markov logic networks. Mach. Learn. **62**(1–2), 107–136 (2006). https://doi.org/10.1007/s10994-006-5833-1
42. Tran, S.D., Davis, L.S.: Event modeling and recognition using Markov logic networks. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 610–623. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88688-4_45
43. Du, Y., Fu, Y., Wang, L.: Skeleton based action recognition with convolutional neural network. In: 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 579–583. IEEE (2015). https://doi.org/10.1109/acpr.2015.7486569