Proceedings of CECNet 2022
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA220524

Real-Time Detection and Early Warning System for Public Security Key Places Based on Human Posture Characteristics

Xiaohui BAI[1], Zhenyu SHOU and Tianhang YUAN
School of Information Technology, Nanjing Forest Police College, Nanjing 210023, China

Abstract. Detention places, as an important part of police's safe of law enforcement, are receiving more and more research attention. We are aiming to apply advanced information technology to the field of detention places and build an early warning prediction model that can improve the security of detention places. This research can accumulate some experience for follow-up technological development in the field of detention places' supervision.

Keywords. Places of detention, security surveillance, risky behavior prediction, personnel control

1. Background Introduction

The need to continuously improve the level of social stability has become increasingly strong. Detention place, as a specific place to detain people who affect social stability, is particularly important in the perspective of security [1]. Only by establishing a set of effective and advanced security system, can we effectively manage the detainees in the daily work of detention place [2].

In response to some abnormal behaviors in Detention place, most of the detention places in China have solved the problem by providing full coverage of surveillance video, which needs to spend the massive manpower, material and financial resources.

2. A review of the literature on security systems for places of detention

2.1. Advances in security algorithms for custodial settings

For the security system of detention places, the core algorithm is mainly pedestrian detection and Human Pose Estimation algorithm.

From the very beginning to 2002, researchers have borrowed and quoted some mature methods in the field of image processing and pattern recognition.

In 2005, Dalal and others [3]. In 2011, Zhu proposed CENTRIST, namely the central transform histogram feature [4]. Several powerful algorithmic frameworks have been proposed so far, including R-CNN, YOLO, and SSD.

[1]Corresponding author: Bai Xiaohui, E-mail: 444340687@qq.com.

In 2013, the research of pose estimation gradually started to shift from the traditional research to the research of deep learning Human Pose Estimation [5]. Alexander Toshev and others further combined DeepPose with CNN. By 2016, Convolutional Pose Machine (CPM) was introduced into Human Pose Estimation algorithms [6]. In the current field of Human Pose Estimation, classical algorithm structures such as CPN, HRNet, etc.

2.2. Selection of algorithms for security systems in detention facilities

2.2.1 Comparison of the advantages of different algorithms

2.2.1.1 Pedestrian detection

The mainstream pedestrian detection are mainly as follows.

RCNN has high precision under specific network model and data set, as shown in Figure 1. However, due to its large number of convolutional neural network computation, RCNN runs slowly and takes up a lot of space to run [7].

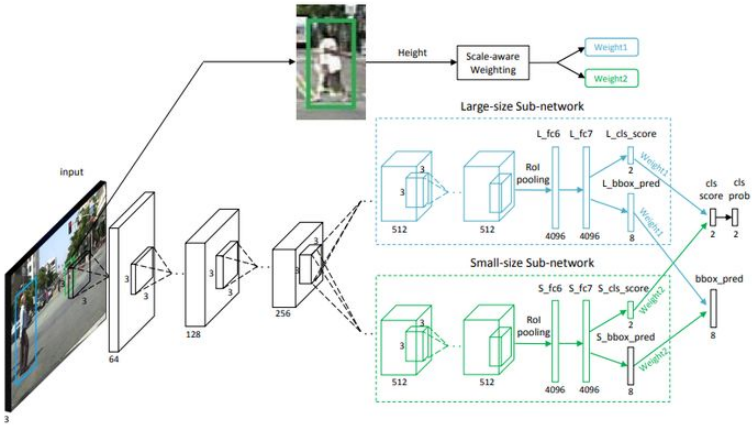

Figure 1. RCNN

SSD balances the advantages and disadvantages of YOLO and RCNN. However, SSD requires manual settings during debugging and cannot be automatically learned, which makes the debugging of SSD model very dependent on experience [8].

YOLO, like SSD, has the characteristics of accuracy and high speed. YOLO has excellent performance in overall detection by end-to-end testing. The YOLO uses CNN networks for target detection, which is very simple and fast [9].

2.2.1.2 Human Pose Estimation

The mainstream of Human Pose Estimation includes MediaPipe, OpenPose, HRNet and so on. MediaPipe has good accuracy and operation speed for Human Pose Estimation, but its gross defect is that cannot recognize the posture of multiple people. Although it can combine with other algorithms to realize the estimation function of multi person posture, the final result lags behind and the accuracy will be greatly reduced [10].

OpenPose integrates the models of mediapipe and posenet, and has high stability. It has good accuracy in most cases and can adapt to different environments by changing

the convolution core weight [11]. However, it is difficult for openpose to avoid misjudgment of actions in videos with complex backgrounds.

The biggest advantage of HRNet is that it can maintain the high resolution of images during operation. However, the operation process of HRNet is very complex and non real-time because it lacks the process of down sample.

2.2.2 Advantages of YOLOv3 and OpenPose algorithms

2.2.2.1 Pedestrian detection

Among the pedestrian detection, we have chosen the YOLOv3.

In this project, the algorithm of pedestrian detection is required to sensitively detect people in detention places. Comparing the mainstream algorithms, RCNN has very high accuracy and very sensitive perception for small targets, but it will consume a lot of time and memory space due to the complexity of its convolutional neural network. Thus, RCNN is not suitable for this project. Similarly, although SSD has a good performance of detection accuracy and running speed, its debugging process not only requires lots of manual settings, but also requires too much experience for the project team. YOLOv3 is much better than SSD variants and comparable tostate-of-the-art models on the APso metric, as shown in Figure 2. Therefore, it is not suitable for this project in all aspects. YOLOv3 is simple, fast, and sensitive to global video detection. Meanwhile, the layering of object detection makes YOLOv3 more effective for small targets, which is more suitable for this project's simple and efficient requirements [12]. Besides, YOLOv3 has good real-time performance, which can well meet the effect of real-time monitoring in this project [13]. So, we selected YOLOv3 to complete the figure detection in this project.

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods							
Faster R-CNN+++ [3]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [6]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [4]	Inception-ResNet-v2 [19]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [18]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
One-stage methods							
YOLOv2 [13]	DarkNet-19 [13]	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [9, 2]	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [2]	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet [7]	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet [7]	ResNeXt-101-FPN	**40.8**	**61.1**	**44.1**	**24.1**	**44.2**	51.2
YOLOv3 608 × 608	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9

Figure 2. Algorithm comparison results.

2.2.2.2 Human Pose Estimation

Among the Human Pose Estimation, we chose OpenPose.

Although the MediaPipe has good accuracy and operation speed for motion recognition, its function can only estimate the pose of a single target and cannot meet the needs of this project. In terms of HRNet, it has the strongest function and is better than the other two algorithms in terms of prediction accuracy, but its function of maintaining image clarity during the operation is not necessary for this project. Besides, HRNet's poor real-time performance during the operation cannot meet the demand of timely alarm.

Therefore, we chose OpenPose. OpenPose as a real-time multi-person keypoint detection library allows simultaneous pose estimation of body, face and limbs [14].

3. Model Building

3.1. Personnel Standard Action Judgment System

The standard action judgment system uses posture evaluation algorithms and key frame action feature extraction to construct an innovative standard technical action data base in public security detention facilities. The core part is divided into 3 units - human key point prediction unit, key point data pre-processing unit, and standard action data database construction unit. The overall flow chart is shown in Figure 3.

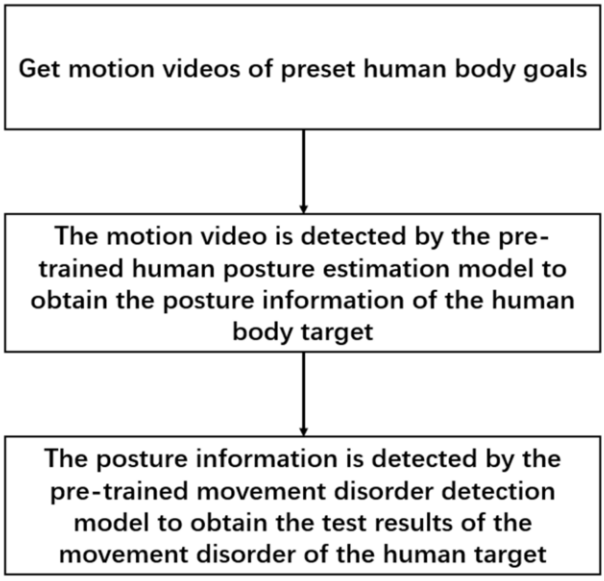

Figure 3. Overall flow chart of the unit.

When the video is input, the system extracts the video frame by frame and analyzes each frame of the video using the target detection algorithm, and then we use the OpenPose algorithm to extract the human bones in each frame, which is the human key point prediction unit.

The 2D or 3D coordinates of each human skeleton in each frame are usually used to represent the skeletal sequence. Previously, action recognition based on skeletal points was done by linking all the joint vectors in each frame into one feature vector. By adding the video to the algorithm and running it, the resulting JSON file is parsed as follows: position information of each body part (x, y, score), each part is an array containing the position information and detection confidence of each body part in the format = x1,y1,c1,x2,y2,c2,... coordinates x and y can be normalized to the intervals [0,1], [-1,1], [0,source size], [0,output size], etc., and the skeletal information of the person can be obtained by concatenating each skeletal point. As shown in Figure 4.

Figure 4. Skeletal Coordinate Chart.

3.2. Character movement recognition

After the character localization and skeleton extraction operations, the system will define and recognize the actions of the annotated joint dynamic data training set by the st-gcn algorithm.

The hierarchical nature of ST-GCN eliminates the need for manual partitioning or traversal of rules. As shown in Figure 5 this not only results in greater expressiveness and higher performance (as shown in our experiments), but also makes it easy to generalize across different environments. Based on the generic GCN formulation, we also designed a new strategy for graph convolution kernels based on image model-inspired research. The main contributions of this work are threefold: 1) We propose ST-GCN, a graph-based approach for dynamic skeletal modeling, which is the first application of a graph-based neural network for this task. 2) We propose several principles for designing convolution kernels in ST-GCN, aiming to meet the specific requirements of skeletal modeling. 3) On two large-scale datasets based on skeletal action recognition, our model requires considerably less manual design and achieves superior performance compared to previously used methods that manually assign partial or traversal rules.

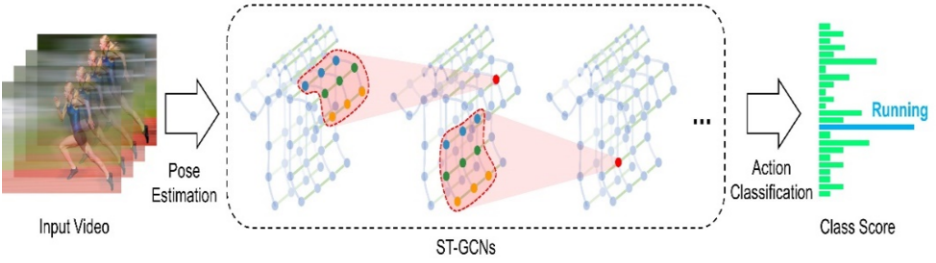

Figure 5. ST-gcn algorithm.

3.3. Motion acquisition subsystem

The motion capture subsystem uses the yolo object tracking detection algorithm and key frame motion feature extraction to build a human motion capture system. When an image

is fed into the yolo network, it is first scaled to a 416 by 416 size. After adding gray tones to the edges of the image to prevent distortion, yolo divides the image into 13 * 13, 26 * 26, and 52 * 522 grids. 52 * 52 grids are used to detect small objects since small features tend to disappear after multiple convolution and compression. 13 * 13 grids are used to detect large objects. Since the cat is a relatively large object, it has a 13 * 13 grid for detection, and each grid point is responsible for the detection of its lower right corner region. If the center point of the object falls in this region, the position of this object is determined by this grid point. yolo is nothing but dividing a picture into different networks, and each grid point is responsible for the prediction of its lower right region. As long as the object's center point falls in this region, this object is determined by this grid point.

When the video is fed into the system, the YOLO algorithm extracts detects and tracks the pedestrians in the video frame by frame. As shown in Figure 6.

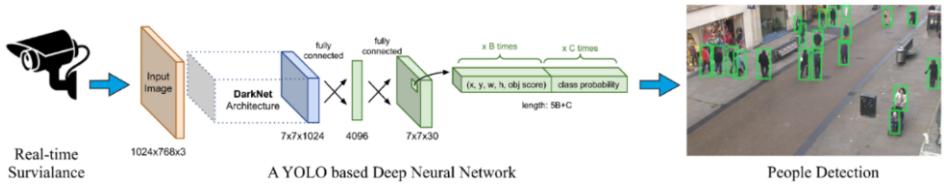

Figure 6. Yolo character extraction.

3.4. RFCOMM Bluetooth Protocol Alarm

Finally, the program will further test and evaluate the action posture of the identified object based on the training set of the annotated joint dynamic data, and then predict the possible abnormal behaviors of the suspect in the detention center, such as self-harm and fall, and define such actions as abnormal actions, if the action is defined as abnormal actions, the program will send out the warning message, and at the same time send the received abnormal warning message to the buzzer. If the action is defined as abnormal action, the program will send out a warning message, and at the same time will receive the abnormal warning information sent to the buzzer, that is, will be based on RFCOMM Bluetooth protocol based on communication programming connected to Bluetooth alarm, timely warning information to warn the police.

4. Conclusion

In this research, we focus on the key areas of social security work and the difficult areas of public security work under the condition of complicated public security situation. It revealed that the public security department does not have an excellent early warning and monitoring model for high-risk abnormal behaviors of specific personnel in detention places. Fortunately, most places of detention in our country are equipped with monitoring and other basic equipment during the process of public security informationization. On the basis of these conditions, we have added the recognition and alarm function of the person's movement to the monitoring to realize the judgment of high-risk abnormal behavior. The whole process is simple and easy to understand, suitable for use in places of detention.

Acknowledgements

This research was support by the 2022 College Students Innovation and Entrepreneurship Training Program (Grant No. 202212213024Z).

References

[1] Grethe Midtlyng. Safety rules in a Norwegian high-security prison. The impact of social interaction between prisoners and officers. Safety Science. 2022 May;149.
[2] Zeping Zhang1,Jing He1, Zhiwei Zhang1. Emotion Recognition Algorithm Based on Panorama-plane Mapping Dataset and VGG16 in Prison Monitoring System. Journal of Physics. Conference Series. 2020; 1627(1): 012010.
[3] Tomoki Watanabe, Satoshi Ito, Kentaro Yokoi. Histograms of oriented gradients for human detection.IPSJ Transactions on Computer Vision and Applications. 2010; 2:39-47.
[4] Wu Jianxin, Rehg James M. CENTRIST: A Visual Descriptor for Scene Categorization. IEEE transactions on pattern analysis and machine intelligence. 2011; 33(8): 1489-1501.
[5] Alexander Toshev, Christian Szegedy. DeepPose. Human Pose Estimation via Deep Neural Networks. IEICE Transactions on Fundamentals of Elect. 2013.
[6] Jian He, Cheng Zhang, Xinlin He, Ruihai Dong. Visual Recognition of traffic police gestures with convolutional pose machine and handcrafted features. Neurocomputing. 2020; 390(prepublish) : 248-259.
[7] Yang Ai-min, Jiang Tian-yu, Han Yang, Li Jie, Li Yi-fan, Liu Chun-yu. Research on application of on-line melting in-situ visual inspection of iron ore powder based on Faster R-CNN. Alexandria Engineering Journal. 2022; 61(11) : 8963-8971.
[8] Gao Xinbiao, Xu Junhua, Luo Chuan, Zhou Jun; Huang Panling, Deng Jianxin. Detection of Lower Body for AGV Based on SSD Algorithm with ResNet. Sensors. 2022; 22(5) : 2008-2008.
[9] Ignacio Martinez-Alpiste, Gelayol Golcarenarenji, Qi Wang, Jose Maria Alcaraz-Calero. A dynamic discarding technique to increase speed and preserve accuracy for YOLOv3. Neural Computing and Applications. 2021; 1-13.
[10] Yasumuro Masanao, Jin'no Kenya. Japanese fingerspelling identification by using MediaPipe. IEICE Communications Society Magazine. 2022; 13(2) : 288-293.
[11] Jiayuan Xing, Jun Zhang, Chenxing Xue. Multi person pose estimation based on improved openpose model. IOP Conference Series: Materials Science and Engineering. 2020; 768(7) : 072071.
[12] Wang Zhihui, Zhu Houying, Jia Xianqing, Bao Yongtang, Wang Changmiao. Surface Defect Detection with Modified Real-Time Detector YOLOv3. Journal of Sensors. 2022.
[13] He Xiaopei, Wang Dianhua, Qu Zhijian. An Improved YOLOv3 Model for Asian Food Image Recognition and Detection. Open Journal of Applied Sciences. 2021; 11(12) : 1287-1306.
[14] Andi W. R. Emanuel, Paulus Mudjihartono, Joanna A. M. Nugraha. Snapshot-Based Human Action Recognition using OpenPose and Deep Learning. IAENG International Journal of Computer Science. 2021; 48(4).