Electronics, Communications and Networks
A.J. Tallón-Ballesteros et al. (Eds.)

doi:10.3233/FAIA231251

Simplified Monocular Ranging Anomalous Behaviour Detection Model for Security Monitoring Applications

Qing LIU[1a], Yu WU [a], Jing NIE [a], Yi LI [a], Xiaohu ZHENG [a], Kang LIU[b] and Ping WANG [c]
[a] *Bijie Power Supply Bureau, Guizhou Power Grid Limited Liability Company, Bijie City, Guizhou Province 551700, China;*
[b] *Power Dispatch Control Centre, Guizhou Power Grid Limited Liability Company, Guiyang City, Guizhou Province 550000, China;*
[c] *Chongqing University of Posts and Telecommunications, Nanan District, Chongqing 400065, China*

Abstract. A simplified monocular ranging anomalous behaviour detection model is proposed to address the problems of cumbersome calibration of the traditional monocular ranging camera and the difficulty of selecting anomalous behaviour detection models. The model uses the YOLO target detection algorithm to obtain the pixel position of the target ROI region in the image coordinate system, and then converts the pixel position to the world coordinate system by using the vertical and horizontal field of view angles of the camera and the camera resolution. Based on the camera mounting height and tilt angle, the distance between the target and the camera vertical point is calculated and compared with the preset defense zone distance and the model switching distance to determine whether the target needs to be detected for abnormal human behaviour and which abnormal human behaviour detection model should be used for detection. In addition to ranging the target without camera calibration, the model also divides the anomaly detection network for two different human features by the image distortion distance, which improves the convergence speed and recognition accuracy of the anomaly detection network.

Keywords. Target detection, monocular ranging, human detection, deep learning

1. Introduction

The ability to detect and recognize abnormal human behaviors from visual data is importance in various domains, including public safety, healthcare, and smart environments. Traditionally, surveillance systems have relied on multiple cameras or complex sensor setups to monitor and analyze human activities[1][2]. The single-camera-based solutions has gained attention due to its cost-effectiveness and ease of deployment. These systems have the potential to provide insights into human behavior, but they face the challenge of estimating distances between the camera and observed individuals, which enables the detection of actions that deviate from typical patterns or are spatially inconsistent with the environment.

[1] Corresponding author: 271070837@qq.com

This paper presents a comprehensive approach to address the problem of monocular distance estimation and abnormal behaviour detection, which is combine the monocular detection model and the proposed distance detection algorithm to estimate the distance. By comparing the detected distance with the set defense zone distance, determine whether and how to use anomaly detection model. This approach considers factors include human image feature aberrations and pose detection models to improve the accuracy of distance estimation and human anomaly detection.

Organization. The rest of the paper is organized as follows: Section 2 provides an overview of related work in monocular human behaviour analysis and distance estimation. In Section 3, we describe our proposed method in detail and highlight the key components and techniques used in it. Section 4 details the experimental steps and performance evaluation methods to demonstrate the feasibility of our approach. Finally, in Section 5, we summarize the paper, discuss its implications and outline potential avenues for future research.

2. Related Work

2.1. Monocular Depth Estimation

Research in monocular depth estimation began with traditional computer vision methods. Early approaches included techniques such as Stereo Vision, Structured Light and Optical Flow using a single camera. These methods have had some success, but there are many limitations in applications. Depth estimation based on deep learning don't depend on special hardware configurations and can estimate depth information from a single image. A series of complex neural networks based on monocular depth estimation such as FCRN [3] have been proposed by researchers. These methods improve the performance of depth estimation by consuming complex hardware setups.

Hardware equipment and environmental limitations often make the acquisition of the depth map itself have a certain error. Lee et al.[4] infer depth by detecting information about the size of target objects within the frame. These methods avoid the errors associated with ground truth depth, but require large datasets and hardware overheads. Other researchers explored weakly supervised and self-supervised learning approaches to reduce the reliance on depth labeling. Casser et al.[5] present a method for self-supervised depth and ego-motion estimation using image sequences. Kuznietsov et al.[6] explore techniques for monocular depth map prediction using labelled depth maps and unlabeled image data to train depth estimation models. The use of different techniques and methods to estimate depth has become the current trend in monocular depth estimation.

2.2. Human Anomaly Detection

Early human anomaly detection methods were based on traditional computer vision techniques such as background modelling, motion detection and feature engineering. With the development of deep learning techniques, deep learning models such as CNN and Recurrent Neural Networks (RNN)[7] are widely used for anomaly detection tasks. In addition to traditional networks using only video data as input, multi-modal approaches combining data from multiple sensors have also been used by researchers[8]. The method improves the detection performance of anomalous behaviors by fusing and learning multiple data features. To address the problem of difficult dataset acquisition

and inaccurate labelling, Pang et al.[9] introduced a pseudo-labelling combined with an iterative optimization mechanism for unsupervised anomaly detection.

Overall, research on human anomaly detection models covers a wide range of techniques and approaches. These researches have applications not only in the field of security surveillance, but also play a role in various fields such as traffic monitoring and medical image analysis. With the continuous development of technology, human anomaly detection models will become more important and accurate in practical applications.

3. Simplified Monocular Ranging Anomalous Behavior Detection Model

3.1. The Overall Framework Analysis

In traditional target ranging, millimeter wave radar has a high loss of wave-guide devices, and laser ranging are susceptible to interference from smoke and dust. The monocular ranging system based on deep learning needs to use different datasets for different projects. The calibration work is complicated, which is not good for the implementation of project works.

The network framework presented in this paper comprises three main components, aiming to address the deficiencies and needs presented above. The first component is the human detection network, which is based on the YOLO algorithm. The second component is dedicated to distance detection and utilizes a simplified distance detection algorithm. The third component is responsible for abnormal behavior detection, and its operational mode and mode selection are controlled by the preceding component. The overall framework is shown in figure 1.

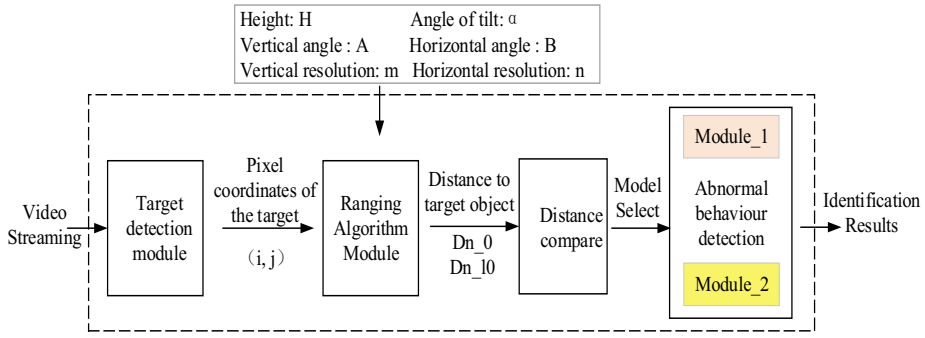

Figure 1. The network overall framework

3.2. Target Detection

YOLO[10] is a popular target detection algorithm whose main idea is to consider the target detection problem as a regression problem instead of the traditional two-step target detection method. Anchor Boxes was introduced to more accurately predict the location of bounding boxes. Each anchor box has a different width and height to accommodate targets of different sizes and shapes. To reduce overlapping bounding boxes, YOLO uses non-extreme value suppression to remove bounding boxes with overlap higher than a threshold. As advantages of fast and accurate, it is widely used well in real-time target detection tasks[11]. The YOLOv5 model is selected as the target

detection algorithm. It's a lightweight version that improves speed by using fewer parameters and smaller model sizes while maintaining comparable accuracy.

3.3. Monocular Distance Estimation

The monocular ranging scheme used in this paper consists of the following steps:

Step 1: As shown in Figure 2a, use the shift method to measure the vertical field of view and horizontal field of view of the camera, to get the maximum field of view in the vertical and horizontal directions, respectively, A, B. The shift method is as follows: place an object in front of the camera, move it to the right and left respectively until it just disappears in the image, and record the position. The angle formed by these two positioning points and the centre of the camera is the horizontal maximum field of view. The vertical field of view is obtained in the same way.

Step 2: As shown in Figure 2b, calculating the distance D_0 between the point E and the pendant point of the camera and the straight-line distance D_{l0} between the point E and the camera, respectively. The specific solution formula is:

$$D_0 = H \Big/ \tan\left(\frac{A}{2} + \alpha\right) \tag{1}$$

$$D_{L0} = H \Big/ \sin\left(\frac{A}{2} + \alpha\right) \tag{2}$$

where A and B are the vertical and horizontal angle ranges of the camera's field of view. α is the angle between the optical axis of the camera and the horizontal line. H is the camera mounting height.

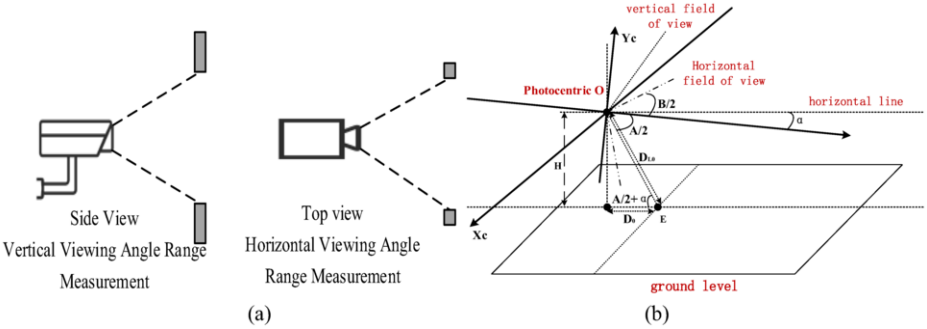

Figure 2. (a) Schematic diagram of field of view measurement and (b) Schematic diagram of field of view measurement.

Step 3: As shown in Figure 3a, derive the coefficient of variation of the field of view per unit pixel in the camera pixel coordinate system corresponding to the camera horizontal and vertical. The formula for the field of view variation coefficient is as follows:

$$\Delta i = \frac{B}{m} * i, i \in [0, \frac{m}{2}] \tag{3}$$

$$\Delta j = \frac{A}{n} * j, j \in [0, n] \tag{4}$$

where m and n are the camera's resolution, respectively. i and j are the pixel values from the midpoint of the lowest row of pixels to the target point. Δi and Δj are the coefficients

of variation of the camera's horizontal and vertical fields of view per pixel.

According to the derived coefficient, the position information of the detected target under the image pixel coordinate system is mapped to the world coordinate system. And then get the distance D_{n_0} between the target point and the pendant point of the camera and get the straight-line distance D_{n_L0} between the target point and the camera. The mapping equations are as follows:

$$D_{\Delta i} = \frac{H}{\cos \alpha} \tan \Delta i \tag{5}$$

$$D_{\Delta j} = H / tan(\frac{A}{2} + \alpha - \Delta j) - D_0 \tag{6}$$

The distance is calculated by the following formula:

$$D_{n\_0} = \sqrt{D_{\Delta i}\text{^}2 + (D_{\Delta j} + D_0)\text{^}2} \tag{7}$$

$$D_{n\_L0} = \sqrt{D_{n\_0}\text{^}2 + H\text{^}2} \tag{8}$$

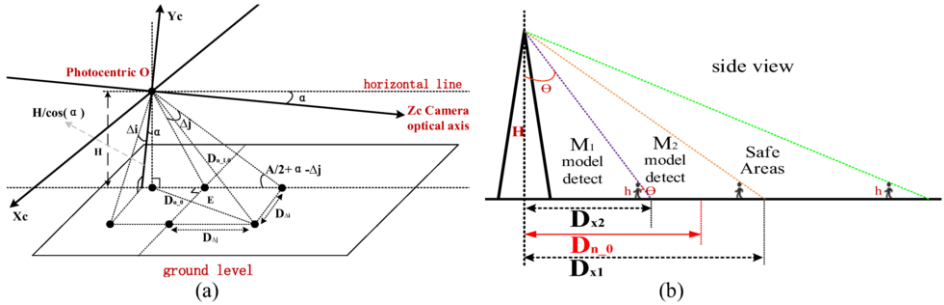

Figure 3. (a) Target Distance Calculation Solution Schematic and (b) Schematic illustration of the zoning of D_{x1} and D_{x2}.

Step 4: As shown in Figure 3b, the obtained distance D_{n_0} is compared with the preset defence zone distance D_{x1} and model switching distance D_{x2}, respectively. If then the target is outside the defence zone, no detection; if $D_{n\_0} \in [D_{x2}, D_{x1}]$, the target ROI region image needs to be input to network M_1 for anomalous detection; if $D_{n\_0} < D_{x2}$, the image needs to be input to network M_2 for anomalous detection.

The distance D_{x2} is the range area where the compression of the human target imaged in the camera is equal to 50%. The formula is as follows:

$$tan\,\theta = \ （h/2）/h = 1/2 \tag{9}$$

$$D_{x2} = \ （H - h）* tan\,\theta \tag{10}$$

where h is the body height.

3.4. Abnormal Behaviour Detection

For better application in real scenarios, this paper uses a lightweight human pose anomaly detection network[12] as Model M_1. This network reduces the risk of disturbing

parameters such as appearance, background, and camera angle affecting the results by focusing the problem on the detection of anomalies in human poses. Furthermore, an anomaly detection framework with adversarial training was trained [13] as Model M_2. The model consists of an object detector, a set of appearance and motion auto-encoders, and a set of classifiers. Since its framework focuses only on object detection, it can be applied to different scenarios.

4. Experimental and Analysis

4.1. Data Set

In this paper, target detection and human abnormal behaviour recognition networks are trained independently. Image data from the person part of the VOC2007dataset is extracted for training the YOLOv5 network. The dataset is divided into training, testing and validation sets, after being transformed into a format suitable for YOLO. Shanghai University of Science and Technology (SUST) campus dataset was used for the training of human anomaly recognition network. The training data contains only normal video samples and the test data contains normal and abnormal videos. Models M1 and M2 are trained using the same dataset.

4.2. Data Preprocessing

In the target detection, the image is adjusted to the input size of the model by scaling and cropping, while data enhancement techniques are used to increase the diversity of the training data and improve the robustness of the model. In the anomalous behaviour recognition phase, the AlphaPose detector is used to detect skeletons in the video frames. At the same time, PoseFlow is used to track the skeleton in the video. Then, each pose sequence is divided into fixed length by sliding window method. Finally, normalization is used to normalize the mean and unit variance of each pose segment to zero.

4.3. Model Evaluation Metrics

The loss function used by YOLOv5 includes target detection loss, confidence loss weights and category loss weights. Mean Square Error (MSE) and Cross-Entropy Loss are typically used. The formulas for calculating MSE and CE are as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right|^2 \tag{11}$$

$$CELoss = -\sum_{i=1}^{N} y_i \log \hat{y}_i \tag{12}$$

where N is the number of samples in the data set, and $\hat{y}_i$ represents the predicted value and is the actual value. In this paper, two networks are trained individually using publicly available datasets. The human anomaly detection component uses Area Under the ROC Curve (AUC) as an assessment metric. To assess the localization of anomalies, we evaluate our model using a region-based detection criterion (RBDC) and a track-based detection criterion (TBDC)

4.4. Experimental Analysis

In this paper, a public dataset and a pre-trained model are used to train the target detection network. The training results are expected to be obtained by framing the person and outputting the pixels of the target box. The midpoint coordinates of the pixels in the bottom row of the target box are selected as the pixel coordinate values of the detected target. By setting the size of the target box, a more reasonable target pixel position can be output. The final training result of this model achieved 95% detection accuracy. The experimental results show that, compared with other models, the YOLOv5 model still has a high detection accuracy while reducing the parameters, which provides data support for the model's subsequent detection.

Model M1 and M2 was evaluated using SHT and SHT HR and the results are presented in Table 1. It could be seen that M1's significantly higher accuracy relative to the other testing models. In addition, the SOTA results of RBDC and the average results of TBDC indicate that M1 and M2 could accurately localize abnormalities.

Table 1. ShanghaiTech Assessment Results

Model	AUC	ShanghaiTech-HR AUC	RBDC	TBDC
Morais et al.[14]	73.4	75.4	--	--
Wang et al.[15]	84.2	84.4	22.3	60.4
M1	85.6	87.3	51.7	82.1
M2	82.5	--	40.9	78.74

The overall framework of the model includes a target detection model, a ranging algorithm, and an anomalous behaviour recognition model. The model training is divided into two parts: target detection and abnormal behaviour recognition. After training, they are connected by ranging algorithm. The range of the defense zone is assumed through the peripheral camera inclination and height. Finally, the integrated model is tested on a public test set. A single M1 and M2 model incorporating the YOLO detection algorithm is used for comparison. The results are presented in Table 2.

Table 2. Model Comparison Results

Model	AUC	RBDC	TBDC
M1	85.7	51.0	81.4
M2	82.7	41.1	76.5
Ours	87.2	50.7	79.8

Experiments show that the model proposed in this paper possesses higher AUC coefficients compared to the one-single detection model. By comparing the RBDC and TBDC indices, the model has a good ability to track the target in the video while maintaining a high judgement accuracy. In addition, the division of the detection defense zone also helps to reduce the false detection rate.

5. Conclusion

The anomalous behaviour detection model proposed in this paper relates the camera resolution to the vertical and horizontal maximum field of view angles of the camera, and constructs a relationship between the coefficient of variation of the camera's per-pixel field of view and the calculation of the target point in the world coordinate system. The target distance is calculated without tedious calibration of the camera. By comparing the target distance with the preset defense zone distance and the model switching distance, we can determine whether the target needs to be detected or not, and which human

anomalous behaviour detection model should be used for detection. In future research, we plan to further analyze and extend the results of this study to more comprehensively assess their applicability and potential applications.

Acknowledgement

This work is supported by the project "Research on Intelligent Monitoring and Communication Technology for Multiple Abnormal Events in High Voltage Transmission Line Corridors" which came from 2022 Guizhou Power Grid Co., Ltd. Bijie Power Supply Bureau under Grant No. 0607002022030101SC00049.

References

[1] Tsai, R. Y., & Hung, Y. P. (2009). A Survey of Multi-Camera Networks. In Advances in Multimedia Modeling (pp. 93-105). Springer.
[2] Gavrila, D. M., Philomin, V., & Vojir, T. (2007). A Multi-Sensor Data Fusion Approach for Pedestrian Detection and Tracking. In Proceedings of the 2007 IEEE Intelligent Vehicles Symposium (pp. 534-539). IEEE.
[3] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, Nassir Navab (2017). Deeper depth prediction with fully convolutional residual networks (oral presentation). IEEE Computer Society.
[4] Lee, Y., Kim, J. (2018). Size to Depth: A New Perspective for Single Image Estimation, CVPR, 2018.
[5] Casser, V., Pirk, S., Mahjourian, R., & Angelova, A. (2019). Unsupervised Monocular Depth and Ego-Motion Learning with Structure and Semantics. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.
[6] Kuznietsov, Y., Stückler, Jrg, & Leibe, B. (2017). Semi-supervised deep learning for monocular depth map prediction. IEEE Computer Society.
[7] Luo, W., Liu, W., & Gao, S. (2017). A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 341-349.
[8] Radu, V., Tong, C., Bhattacharya, S., Lane, N. D., Mascolo, C., Marina, M. K., & Kawsar, F. (2017). Multimodal Deep Learning for Activity and Context Recognition. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1(4), 157, 1–27.
[9] Pang, G., Yan, C., Shen, C., van den Hengel, A., & Bai, X. (2020). Self-Trained Deep Ordinal Regression for End-to-End Video Anomaly Detection. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 12170-12179). IEEE.
[10] Yanhua, S., Duo, Z., Hongyu, C., Xiaoqiang, Z., & Yunbo, RAO. (2022) A Review of YOLO Object Detection Based on Deep Learning. Journal of Electronics & Information Technology, 44(10) pp. 3697-3708.
[11] Hongmin, Z., Pingping, L., Xiaobing, F., & Hong, L. (2022). Improved yolov3 network model for human abnormal behaviour detection. Computer Science, 49(4), 6.
[12] Hirschorn, O., & Avidan, S. (2022). Normalizing Flows for Human Pose Anomaly Detection. arXiv preprint arXiv:2211.10946.
[13] Georgescu, M. I., Ionescu, R. T., Khan, F. S., Popescu, M., & Shah, M. (2021). A background-agnostic framework with adversarial training for abnormal event detection in video. IEEE Transactions on Software Engineering.
[14] Morais, R., Le, V., Tran, T., Saha, B., Mansour, M., & Venkatesh, S. (2019). Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1, 2, 3, 6.
[15] Wang, G., Wang, Y., Qin, J., Zhang, D., Bao, X., & Huang, D. (2022). Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In European Conference on Computer Vision (ECCV), 1, 2, 3, 6, 7.