Human Trespass Detection Based on Lightweight YOLO-v5 and RNN in Restricted Area

Ligen Tian[1, †, *] and Yaoqing Wang[2, †]

[1]School of Science, China University of Mining and Technology, Beijing, China

[2]School of Computer Network Security, Chengdu University of Technology, Chengdu, China

*Corresponding author: 1510730110@student.cumtb.edu.cn

† These authors contributed equally

Abstract. Trespassing endangers the security of individuals and property, disrupts social order, undermines social trust and increases the number of social groups used to maintain social order. In this paper, a new contribution as a method to combat trespassing which involves the monitoring of human behavior for prediction is presented. This method includes two parts: image and text description. In this work we investigate lightweight human behavior detection models based on YOLO-v5 and RNN. We use the same dataset for different models and study various model metrics (e.g., model accuracy and running speed) to compare the performance of different models. For image and video, we used pruning algorithm to lightweight the YOLO-v5 model while ensuring accuracy. For text description, we used different Image-Caption (RNN and CLIP) models to describe human behavior. Finally, corresponding validation experiments were implemented to validate the method proposed in this paper.

Keywords: YOLO-v5; RNN; Trespassing; Deep learning; Behavior description.

1. Introduction

Trespassing is the unauthorized or unauthorized entry onto another person's private property or home, which is unethical, illegal and can have serious consequences. Security personnel need to be posted in key areas to verify the identity of those entering and leaving and, most importantly, to observe the behavior of those entering and leaving. However, regular patrols of potential break-in locations are not feasible due to the cost of personnel. As a result, the need to design automated intrusion detection and early warning prediction techniques using machine learning was raised.

To address this need, we propose a lightweight model based on YOLO-v5 and RNN to implement real-time human behavioral and expression recognition. This paper details the theory and techniques of the lightweight model based on YOLO-v5 and RNN, including target detection, sequence modelling and feature extraction. Through experimental analysis, the paper demonstrates the high accuracy and speed of YOLO-v5 and RNN-based lightweight models for human behavior. Finally, the paper discusses future directions and challenges for lightweight models based on YOLO-v5 and RNN in the field of human behavior and expression recognition, including the enrichment and diversity of datasets, model optimization and performance enhancement, and suggests possible solutions.

2. Introduction and analysis of relevant technical components

2.1. Technical concepts related to computer vision

Image classification: Image classification is the process of assigning an image to a predefined category. Typically, image classification algorithms use machine learning techniques to learn the features of different categories and use these features to classify new images. The main steps of image classification include data preprocessing, feature extraction, classification model training, and model evaluation.

Target detection: Target detection is an important task in the field of computer vision, which aims to identify the location of an object in an image or video and to classify it at the same time. Unlike image classification, target detection requires determining the location and size of an object and classifying it. The main steps of target detection are candidate region generation, object feature extraction, object classification, region filtering, and object localization.

Image Captioning: Image Captioning is a technique for converting images into natural language descriptions. Image Captioning is a combination of computer vision and natural language processing, which requires deep learning techniques [1]. The main steps involved in Image Captioning are: image feature extraction, sequence generation, model training, and model evaluation.

Text embedding: text embedding is the process of converting text into vector form, and this vector representation can better capture the semantic information of the text. Text embedding techniques are widely used in the field of Natural Language Processing (NLP). The advantage of text embedding technique is that it can convert textual information into digital form, which can be more easily processed and analyzed in computers. Also, text embedding techniques can be used to improve processing efficiency and accuracy by methods such as dimensionality reduction [2].

2.2. Technical concepts related to neural networks

Deep Learning: Deep learning mimics the structure of human brain neural networks and performs feature extraction and pattern recognition through multiple layers of neurons to achieve advanced processing and analysis of data such as images, videos, and speech. Based on a given logical structure, it analyzes data and thus draws conclusions similar to humans. It uses multilayer neural network algorithms thus understanding the given data [3].

RNN: Recurrent Neural Network is a deep learning based neural network model for processing sequential data. The main advantage of RNN is that it can take into account the dependencies between sequence data and can better handle temporal data. the basic structure of RNN includes an input layer, a hidden layer and an output layer, in which the neurons in the hidden layer realize the transmission and memory of information through recurrent connections. Common RNNs include basic RNN, LSTM and GRU, etc [4]. Suitable models can be selected for modeling and training according to different tasks.

YOLO-v5: The YOLO algorithm uses a single neural network to perform the entire target detection task, so it is much faster than other target detection algorithms. However, because the YOLO algorithm segments the image into larger grids, it is not accurate enough for small object detection [5]. The YOLO-v5 algorithm, a deep learning-based target detection algorithm, is the latest version of the YOLO family. Its algorithm structure mainly contains CSPNet as the backbone network, SPPNet as the neck network, the FPN structure of YOLOv3 as the head network, and CIoU Loss as the loss function. the network structure of YOLOv5 is a single-scale detection network which uses different convolutional kernel sizes in order to capture features at different scales. In addition, YOLOv5 uses a number of data enhancement techniques such as random flipping, random scaling and random brightness, which allow for a more robust model to be trained.

Lightweighting: Lightweighting refers to reducing the size and computation of a model through a series of optimization methods, thus improving the speed and deployment efficiency of the model while ensuring its performance. In the field of deep learning, lightweighting has become increasingly important for application scenarios that need to run on resource-constrained devices due to the high model complexity and computational effort. Common lightweighting methods include model pruning, quantization, distillation, etc. These methods can reduce model size and computation by removing unnecessary parameters, reducing model precision, and model compression [6].

3. Methodology

3.1. Overall framework

In our method, a camera is used to capture behavioral and location information, a YOLO-v5 model is trained to detect the human position using the VOC dataset and an RNN model is trained to describe the behavior using the flickr30k dataset. Then the human location is detected using YOLO-v5, the region of interest is extracted, the NMS algorithm distinguishes the target, and the CLIP model classifies the behavior. The frames were also converted to grey scale/color images and scaled to the same size. The behavior is then classified using the CLIP model, the behavioral relationships are described using the RNN model and the cross-entropy loss function trains both models. Human behavior videos were captured, the YOLO-v5 model was trained, and the annotated dataset was used to detect human locations while the RNN model was trained to generate human behavior descriptions. The dataset was divided into a training set, a test set and a validation set. the YOLO-v5 model used the VOC human dataset and the RNN model used the Flickr 30k dataset. To reduce the complexity and computational effort of the model, we pruned the model using the FPGM pruning algorithm. Specifically, we pruned the redundant parameters and connections in the YOLO-v5 model and the RNN model and used the pruned models for testing. The model performance was evaluated using the test set, CLIP was classified using precision, recall, and F1 score, RNN was described using BLEU, ROUGE, and the number of target detection frames was used to determine if the number of people corresponded. Finally, visualization tools were used to present, interpret and analyze the model results, compare model performance and scale before and after pruning, and evaluate the effect of pruning.

3.2. Improvement of algorithm and model

YOLO-v5 is a target detection algorithm which is an improvement and optimization of YOLO-v4. The YOLO family of algorithms has become the most popular algorithm in the field of target detection due to its high performance. The latest generation of YOLO-v5 has a weight file of only 28MB, which is well suited as an initial model [7]. YOLO-v5 improves the accuracy and robustness of the model while maintaining its speed and lightness by optimizing the model structure and adopting new techniques. As such, it is a good method for detecting human behavior and expressions. We prune the model using the FPGM pruning algorithm, a model pruning algorithm based on the L1 paradigm that reduces the size and computational effort of the model by pruning out redundant parameters, while maintaining the accuracy of the model. The main idea is to select the parameters to be retained by calculating their importance and to rank the parameters using global information. Specifically, the FPGM algorithm is divided into two phases: global importance assessment and pruning. In the global importance assessment phase, the FPGM algorithm uses a benchmark model to calculate and rank the importance of each parameter. The calculation of importance is usually based on the L1 paradigm, as the L1 paradigm allows for the deletion of unimportant parameters by reducing their values to near zero. In the pruning phase, the FPGM algorithm prunes the less important parameters and adjusts the size of the remaining parameters according to their importance [8]. Process of whole system is shown in Figure 1.

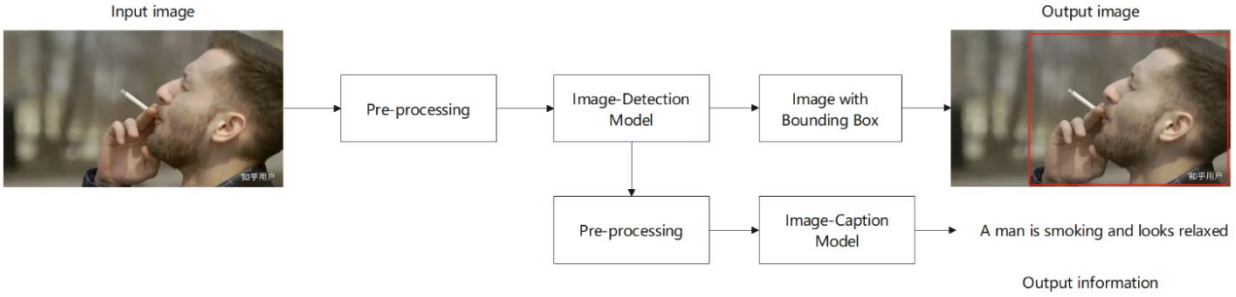

Figure 1. Process of whole system [9]

3.3. Collection of the required datasets

We collected the dataset from Kaggle, which in combination with the clip model will be used to be responsible for behavior detection. The personnel behavior status in this dataset mainly includes: walking, standing, waving, hugging, interacting, and running. The number of training samples is 29000; the number of testing samples is 1014; and the number of validation samples is 1000.

3.4. Pre-processing

In this study, the preprocessing has the following steps:

1. Data collection: Collect data sets of various human behaviors and expressions.
2. Image processing: Perform image preprocessing on image sequences, including image enhancement, denoising, size normalization, etc.
3. Target detection: Perform target detection using YOLO-v5 to identify targets of human behaviors and expressions.
4. Sequence construction: Construct the target detection results into sequences for the next step of RNN model training.
5. Feature extraction: Use convolutional neural network to extract feature vectors of each target as the input of RNN model.
6. RNN model training: Use RNN model to train the input feature vector sequence to recognize human behaviors and expressions.
7. Model optimization: Optimize the trained model, including parameter adjustment, model structure optimization, etc.

In the preprocessing process of this paper, YOLO-v5 is used for target detection to identify human behavior and expression targets in each frame of the image, and then the recognition results are constructed as sequences to facilitate subsequent RNN model training. the RNN model is then used to process the target sequences, extract features and perform classification. Such a pre-processing process can achieve lightweight human behavior and expression recognition with high efficiency and accuracy.

3.5. Split the data

According to the collection of the required datasets, we divided the dataset into three groups: 93.5% of the training set, 3.2% of the test set, and 3.3% of the validation set.

4. Experimental results and analysis

4.1. Experiment setting

The experimental environment is Ubuntu 18.04 on Intel I9-12900H CPU and NVIDIA RTX3080 GPU machine. We use the image and label information extracted from the VOC2012 and Flickr30K dataset. VOC2012 only contains images and labels of the type of version. Therefore, the Yolov 5 section uses 8000 images and corresponding labels as the dataset. Flickr30K, the core of this dataset consists of two things: the images and the language used to describe them [10]. Each image is accompanied by five descriptions, all of which have a similar meaning.

4.2. YOLO-v5 light-weight experiment

The current YOLO-v5 uses the FPGM lightweight pruning algorithm, where the pruning rate is the proportion of convolutional kernels removed from each convolutional layer. Using the VOC dataset, the model was tested with FPGM pruning rates of 12.5%, 25% and 50%, and the corresponding experiment results are shown in bellow. The experimental results obtained after 250 epochs of training with the parameter set at 12.5% are shown in Table1:

Table 1. Performance of pruned model with the parameter set at 12.5%

States	Original model size	Number participants	Calculated volume	Loss precision
Lightweight front	25.1%	12322812	16.9	0.637
Lightweight after	18.8%	9759976	13.7	0.622
Percentage decrease	25%	21.43%	18.93%	0.15

The experimental results obtained after 250 epochs of training with the parameter set at 25% are shown in Table2:

Table 2. Performance of pruned model with the parameter set at 25%

States	Original model size	Number participants	Calculated volume	Loss precision
Lightweight front	25.1	12322812	16.9	0.637
Lightweight after	14,3	7414388	10.8	0.624
Percentage decrease	43.03%	40.31%	36.09%	0.13

The experimental results obtained after 250 epochs of training with the parameter set at 50% are shown in Table3:

Table 3. Performance of pruned model with the parameter set at 50%

States	Original model size	Number participants	Calculated volume	Loss precision
Lightweight front	25.1	12322812	16.9	0.637
Lightweight after	7.3	3674956	6.0	0.604
Percentage decrease	70.92%	70.18%	60.95%	0.33

As the pruning amplitude of the model increases, the size, parameter quantity, and computational complexity of the model will decrease large pruning can cause a significant loss of accuracy. However, from the performance results, the trained epochs can be appropriately improved, because theoretically, 25% pruning will definitely lose more accuracy than 12.5% pruning, and 50% pruning should not lose so much accuracy.

4.3. Experiment and analysis of CLIP & RNN model

The RNN part uses the N VS M structure of the RNN structure, and the model contains the encoder-decoder structure, which is essentially similar to the translation task, except that the images are translated into text as much as possible, and the training uses the flickr 30k dataset.

The RNN model is a generative model that tends to generate corresponding textual descriptions based on images; the CLIP model is a discriminative-like model that tends to 'calculate' the closest distance to a pre-defined textual description for images. model, which tends to compare existing images with pre-defined text descriptions, 'computing' the closest text description for the image, and then using this as the corresponding text description. The CLIP model is a discriminative model that tends to compare existing images with pre-defined textual descriptions, 'computing' the closest textual description for the image and using this as the corresponding textual description. They have something in common in that they both use the encoder- Decoder model structure, but the difference is that RNN uses the full structure for the transformation, whereas CLIP uses encoder similarly to encode the images and text are then compared.

For facial expression descriptions, we use a total of 3173 images as test data, of which 879 are labeled Happy, 1688 are labeled Sad, and 626 are normal. For behavior description, we use 1244 images containing the tags as the test set: walking has 91 images, running has 64, boxing has 193 images, smoking has 805, riding has 34 and sitting has 57. Experimental comparisons were conducted using RNN and CLIP on the above datasets, and the results are shown in Table4.

Table 4. Comparison between RNN and CLIP

	RNN	**CLIP**
Expression accuracy	78.88%	94.61%
Behavior accuracy	56.52%	92.44%

From the experimental results, it can be seen that the ready-made CLIP model performs much better than the trained RNN model, for the following reasons: The number of training datasets in different categories is uneven; The epochs trained on the model are relatively few, and the model has not been fully trained.

5. Conclusion

In this paper, we integrate the CLIP model and the YOLO-v5 model into a lightweight human behavior model as a means to determine the purpose of an entrant. Our experimental results show that the integrated model can effectively identify human behaviors and make accurate inferences about the intent of the entrants. The model is highly accurate and scalable, and can be widely used in various practical application scenarios, such as security surveillance and smart home. This study provides useful references and lessons for further exploration of new technologies and applications in the field of computer vision. In addition, our study shows that both the YOLO-v5 model and the RNN model we use are lightweight models that do not require much computational resources and time, and thus can process video frames quickly. This allows our model to perform real-time detection and description with lower computational resources.

However, we should also note some limitations of the model. First, the recognition capability of the model is still limited by the quality and quantity of data, and the performance of the model may be affected if there is not enough high-quality data for training. Second, our model may be slightly inferior in speed relative to some state-of-the-art models. Since we are using a lightweight model, it may not be able to handle large-scale video data and complex behavioral scenarios.

References

[1] Hossain M D Z, Sohel F, Shiratuddin M F, et al. A comprehensive survey of deep learning for image captioning [J]. ACM Computing Surveys (CsUR), 2019, 51(6): 1-36.

[2] Tang J, Qu M, Mei Q. Pte: Predictive text embedding through large-scale heterogeneous text networks[C]//Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 2015: 1165-1174.

[3] Sharadhi, A.K. et al. (2022). “Face mask recogniser using image processing and computer vision approach”. In: Global Transitions Proceedings, pp. 67–73.

[4] Liu W, Guo P, Ye L. A low-delay lightweight recurrent neural network (LLRNN) for rotating machinery fault diagnosis[J]. Sensors, 2019, 19(14): 3109.

[5] Wu, Wentong, et al. "Application of local fully Convolutional Neural Network combined with YOLO v5 algorithm in small target detection of remote sensing image." PloS one 16.10 (2021): e0259283.

[6] Ma M, Wang J, Yu Z. Differentiable Network Pruning via Polarization of Probabilistic Channelwise Soft Masks[J]. Computational Intelligence and Neuroscience, 2022, 2022.

[7] Lv H, Yan H, Liu K, et al. Yolov5-ac: Attention

[8] Mechanism-based lightweight yolov5 for track pedestrian detection[J]. Sensors, 2022, 22(15): 5903.

[9] Jiang X, Wang N, Xin J, et al. Learning lightweight super-resolution networks with weight pruning[J]. Neural Networks, 2021, 144: 21-32.

[10] Tadas Baltrusaitis, Chaitanya Ahuja, Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. [cs.LG], 2017.

[11] Plummer B A, Wang L, Cervantes C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2641-2649.