International Journal of Computational Intelligence Systems and Applications. Volume 12, Issue 2, (2025) 1-7
http://www.cayley-nielson.com/ijcisa/

Design of Human Pose and Action Recognition System Based on YOLOv8

Li Jiayao[1], Cui Dongyan[1,*]

[1] School of Artificial Intelligence, North China University of Science and Technology, Tangshan, Hebei, China

*Corresponding author

E-mail: *cdy_xxz@163.com*

Abstract

With the surge in demand for accurate human pose and action recognition in fields such as intelligent surveillance and human-computer interaction (HCI), traditional methods struggle to balance accuracy and real-time performance in complex scenarios. YOLOv8, with its exceptional detection speed and accuracy, has emerged as an ideal choice for building related systems. This study develops an efficient and accurate human pose and action recognition system based on YOLOv8, catering to the demands of high-precision and real-time recognition across multiple fields. The system construction encompasses four modules: data collection, preprocessing, model construction, inference and application. The data collection module acquires a high-quality and diverse datasets, laying a solid foundation for model training. The preprocessing module enhances the model's generalization ability through data cleaning, annotation, and augmentation. The model construction module optimizes network parameters and determines training strategies based on YOLOv8, addressing the issues of overfitting and slow inference. The inference and application module deploys the trained model in real-world scenarios, enabling real-time human pose and action recognition. Experimental results show that, in comparison with traditional pose recognition methods, the YOLOv8-based human pose and action recognition system exhibits stronger robustness and higher accuracy when dealing with complex scenarios and multi-person pose recognition.

Keywords: Deep Learning; Human Pose Behavior Recognition; Object Detection; System Design.

1. Introduction

As a key research direction in computer vision, Human Pose Recognition has attracted considerable attention with the development of artificial intelligence (AI) and computer technologies. European and American countries, by virtue of their scientific research capabilities and financial advantages, have long been the forefront of this field, while top international journals and conferences serve as the core carriers of research results. In terms of journals, publications such as IJCV (International Journal of Computer Vision), CVIU (Computer Vision and Image Understanding), PAMI (IEEE Transactions on Pattern Analysis and Machine Intelligence), and IVC (Image and Vision Computing); in terms of conferences, events like CVPR (IEEE Computer Society Conference on Computer Vision and Pattern Recognition), ICPR(International Conference on Pattern Recognition), ECCV (European Conference on Computer Vision), and ICIP (International Conference on Image Processing) — all of these publish a large number of the latest research findings related to Human Pose Recognition, covering key methodologies such as deep learning, multi-camera technology, and multi-information fusion.

The landscape of international representative research findings is rich: Alexander Toshev et al. published the paper titled DeepPose: Human Pose Estimation via Deep Neural Networks, which proposes a human pose estimation method based on Deep Neural Networks (DNN). Leveraging the latest advances in deep learning, this method employs a sequence of such DNN regressors to achieve high-precision pose estimation and reasons about pose in a holistic manner [1].The team led by Sven Kreiss published the paper titled PifPaf: Composite Fields for Human Pose Estimation, which proposes a novel bottom-up approach for multi-person 2D human pose estimation. This method is particularly suitable for urban transportation scenarios such as autonomous driving and delivery robots [2].The team led by Vasileios Belagiannis published the paper titled 3D Pictorial Structures for Multiple Human Pose Estimation. This work resolves detection ambiguities by triangulating body joints in paired camera views, introduces a novel 3D pictorial structure model, and infers 3D human configurations from a simplified state space [3].

In recent years, domestic research has yielded remarkable achievements [4-7]. In 2020, Ma Le published Action-Based Human Target Recognition and Pose Estimation Based on Deep Learning, in which a three-stage framework named "DN-2DPN-3DPN" was proposed for moving humans — the DN Network (Detection Network) detects human bounding boxes, the 2DPN Network (2D Pose Estimation Network) estimates 2D poses, and the 3DPN Network (3D Pose Estimation Network) obtains 3D poses [8].In 2021, He Peng published 3D Human Pose Recognition Based on Deep Learning, in which three methods were proposed: 3D human static pose recognition based on multi-view convolutional neural networks, 3D static pose retrieval based on Siamese Networks, and 3D dynamic pose recognition based on LSTM — addressing the relevant core issues [9].In the same year, Lian Jing-xiang published Research on Human Pose Recognition Based on Deep Learning. Aiming at joint prediction errors and the dilemmas of the top-down approach in crowded scenes, this work proposed a high-r-esolution feature-based pose recognition method and a lightweight spatial-channel attention model [10].

In 2022, Yang Haihong published Video-Based Human Action Recognition Method Based on Deep Learning and Pose-Driven Feature Integration, in which a dynamic fusion model was proposed. This model can dynamically combine appearance and pose features according to the reliability of pose information [11]. Also in 2022, Yang Guangyao published Design and Implementation of a Human Pose Recognition System Based on Deep Learning. In this work, the OpenPose algorithm was analyzed and implemented, with the study noting that as a mainstream algorithm, OpenPose balances high recognition accuracy and speed, meeting the needs of commercial development [12].

2. Design and Implementation of a Human Pose and Action Recognition System Based on YOLOv8

2.1. Dataset Preparation

Given that deep learning-based human action recognition systems require extensive human action data to support recognition performance, multi-source collection was conducted for the image dataset: images were gathered from platforms including Aigei.com (a Chinese resource platform for creative materials), Baidu Search Engine, Firefox Browser, and Roboflow (a specialized datasets platform). Additionally, a Python web crawler was used to collect supplementary images from selected websites. After collection, the images were integrated into a human action image dataset named "images," which was stored in a dedicated folder. The final dataset comprises 1,750 images, with 800 sourced from Roboflow and 900 from other aforementioned platforms. After collection, the dataset was divided into a training set and a validation set to lay the foundation for subsequent model training.

Data annotation involves adding information such as labels and bounding boxes to samples, enabling

deep learning algorithms to interpret data. The quality of annotation directly impacts model performance, and manual involvement is required to ensure accuracy and consistency. Since the YOLO model will be used subsequently, the LabelImg annotation tool was adopted to annotate images into the txt format recognizable by YOLO. LabelImg supports multiple platforms including Windows, Linux, and macOS. As an open-source tool, it is compatible with formats such as Pascal VOC, YOLO, and TFRecord. It allows users to draw bounding boxes around target objects in images to generate annotation files and supports customized extensions as needed, providing supervised learning training data for the model.

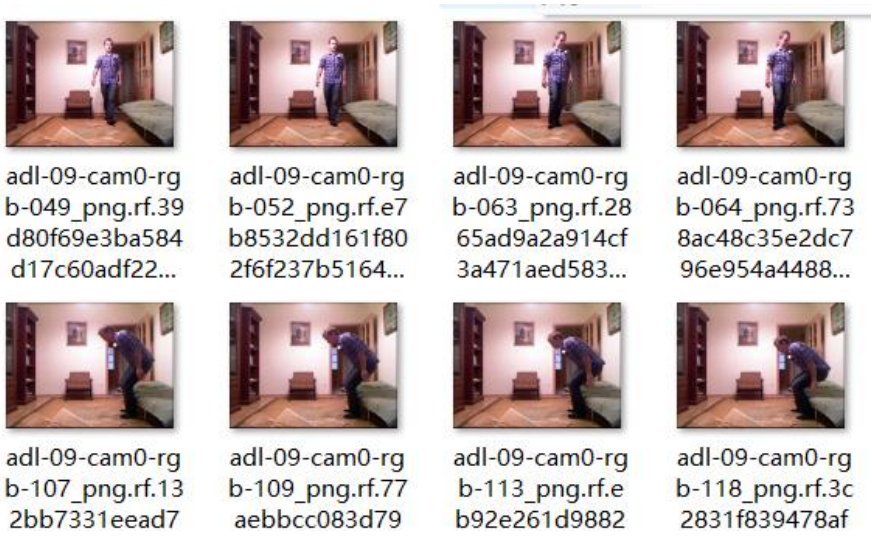

Fig.1 Datasets

2.2. Model Construction

Given its pivotal role in the field of object detection, YOLOv8 boasts advantages such as high performance, lightweight design, strong usability, and rich functionality. Its architecture comprises a lightweight backbone network, a feature fusion layer, and a detection head. Additionally, YOLOv8 optimizes parameters by improving the loss function and balances the importance of different detection tasks.

Figure 2 presents a comparison chart of YOLOv8 and previous versions of the YOLO series. From the comparison, the relative advantages of YOLOv8 can be observed.

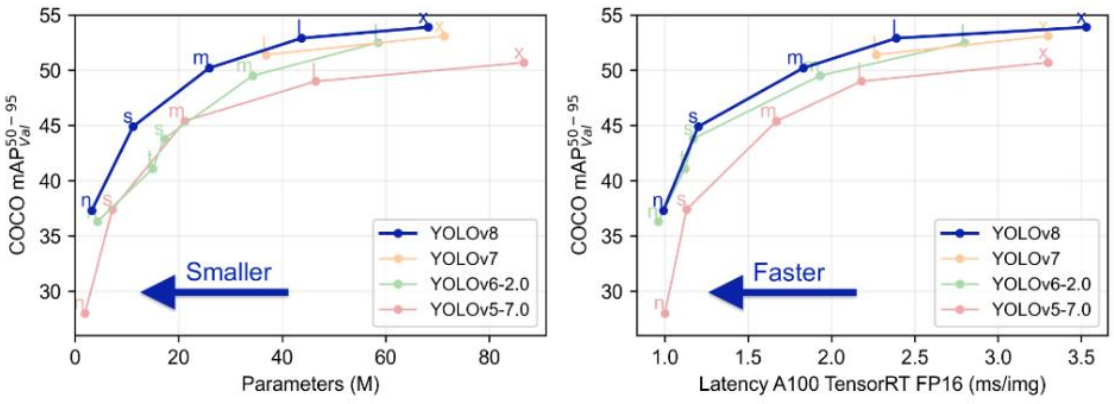

Fig.2 Model Structure Comparison Diagram

YOLOv8 adopts an innovative backbone network and neck structure, integrating the advantages of Transformers and Convolutional Neural Networks (CNN) to effectively enhance its image feature extraction capability. Through constructing a cross-scale feature interaction mechanism and a dynamic candidate box generation strategy, this detection framework significantly improves the model's perceptual robustness in complex environments. Specifically, YOLOv8 employs an Anchor-Free Decoupled Head to replace the traditional anchor-dependent paradigm. By separating the feature representation spaces for classification and regression tasks, it effectively addresses the issues of dimension preset bias and hyper parameter sensitivity in traditional detection methods. This innovation in topological structure endows the detector with superior spatial perception capability, exhibiting significant advantages especially when processing irregular geometric targets. The structure of the detection head is shown in Figure 3.

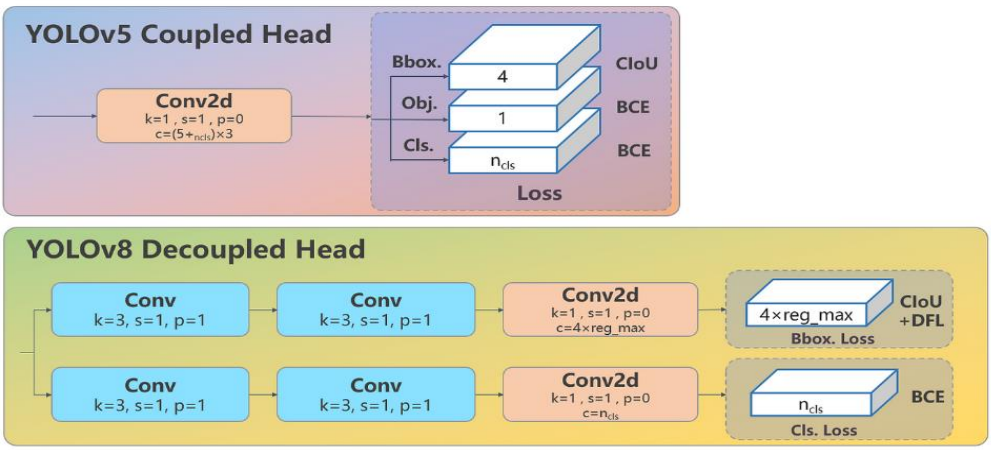

Fig.3 Detection Head Structure Diagram

In summary, with its strengths and innovations such as high efficiency, advanced network structure, innovative detection head design, as well as wide applicability and flexibility, YOLOv8 has achieved significant progress in the field of object detection.

3. Testing and Result Analysis of the Human Pose and Action Recognition System

The design of this system enables it to accurately recognize various types of human actions, including standing, sitting, and falling. Test results indicate good confidence levels, fast inference speed, and clear positional information.

From the recognition results in Figures 4-5, the system exhibits a confidence level of 86%-94% for walking and standing detection, respectively, with relatively large fluctuations in confidence levels. Additionally, it features fast inference speed and high reliability in target detection.

From the recognition results in Figures 6-7, the system exhibits a confidence level of approximately 93% for sitting action detection. Additionally, it features fast inference speed and relatively high reliability in target detection.

By evaluating the result files of the trained human action detection model, we can understand the model's performance and identify three key loss metrics: Localization Loss (box loss) measures the error between predicted boxes and annotated boxes using Generalized Intersection over Union (GIoU). The smaller the GIoU value, the more accurate the localization; Classification Loss (cls_loss) calculates the

correctness of the category matching between anchor boxes and their corresponding annotated categories. The smaller the value, the more accurate the classification; Dynamic Feature Loss (DFLLoss) is a loss function used to regress the distance between predicted boxes and target boxes; During the loss calculation process, target boxes need to be scaled to the feature map scale—specifically, divided by the corresponding stride. Subsequently, the Complete Intersection over Union Loss (CIoULoss) is computed between these scaled target boxes and the predicted bounding boxes. Simultaneously, the DFLLoss is calculated for regression by using the distances from the center points of the predicted anchors to each edge. This process is part of the YOLOv8 training pipeline. By calculating DFLLoss, the position of predicted boxes can be adjusted more accurately, thereby improving the accuracy of target detection. The training results of this project are as follows: The Precision-Recall (PR) curve is typically used to illustrate the relationship between Precision and Recall, and the PR curve of the training results in this work is presented below. Mean Average Precision (mAP) refers to the area enclosed by plotting Precision on one axis and Recall on the other. It can be observed that the model in this work yields a favorable result for mAP@0.5. Figure 8 presents the evaluation and analysis chart of this model.

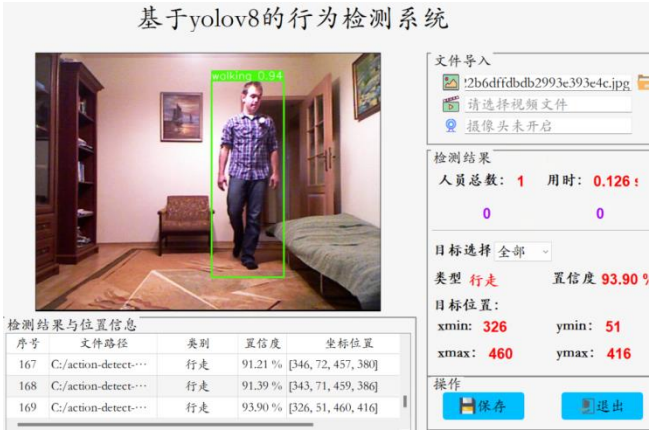

Fig.4 Experimental Test Result Graph (Walking)

Fig.5 Experimental Test Result Graph (Standing)

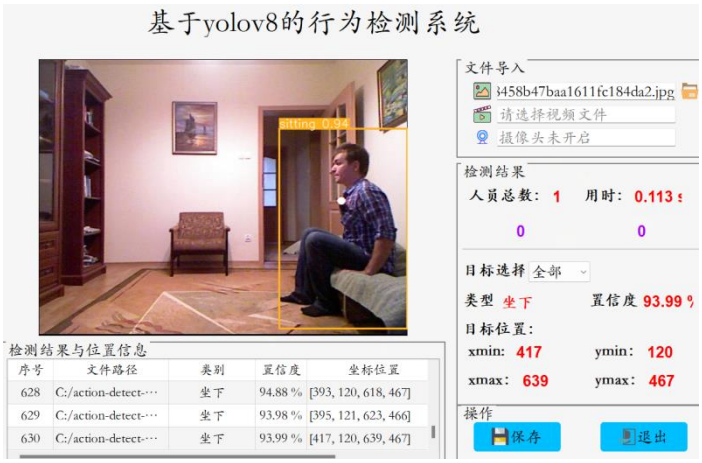

Fig.6 Experimental Test Result Graph (Sitting)

Fig.7 Experimental Test Result Graph (Sitting)

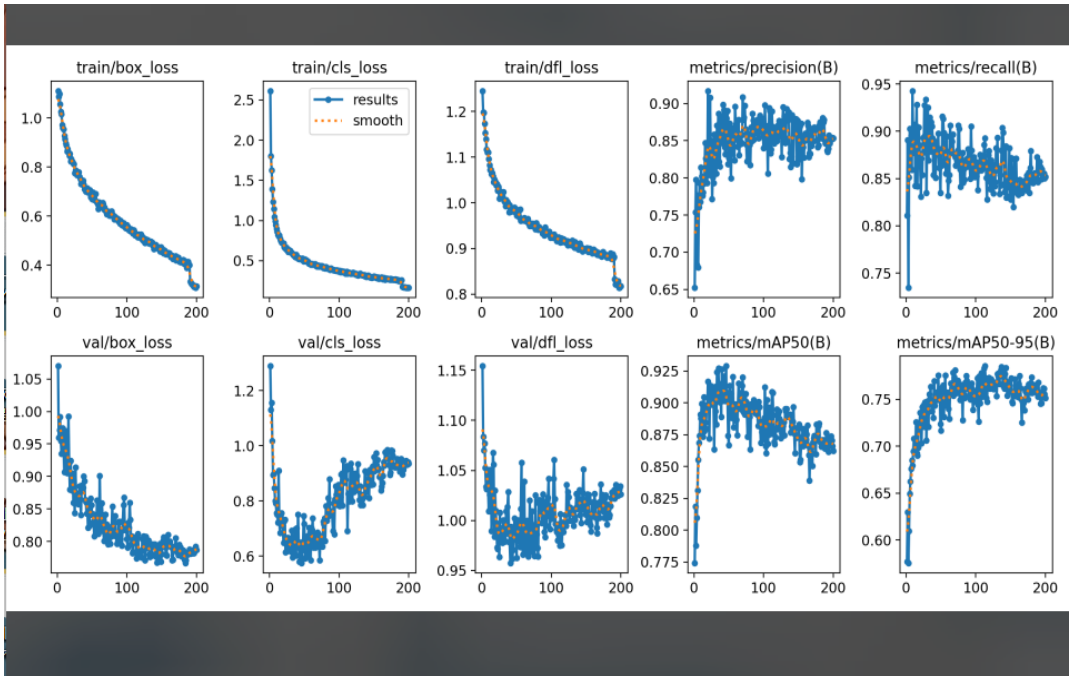

Fig.8 Model Evaluation and Analysis Diagram

Experimental results verify that the adopted deep learning-based human pose recognition model exhibits favorable performance in human pose recognition tasks. It not only achieves a significant improvement in recognition accuracy but also meets real-time inference requirements, thereby providing an effective solution for practical applications.

4. Conclusion

This work constructs a human pose recognition system by adopting the YOLOv8 model. This model outperforms traditional methods and previous YOLO versions in both detection speed and accuracy, enabling it to meet the requirements of scenarios demanding high real-time performance. In the research, deep learning-based recognition approaches were explored: specific efforts included designing the combination of convolutional layers and parameter settings to extract image features; enhancing the model's generalization ability through datasets optimization and preprocessing; and improving recognition accuracy by integrating strategies such as selecting appropriate loss functions, optimizers, and adjusting the learning rate. As a result, the model demonstrates enhanced robustness and accuracy in complex scenarios and multi-person recognition tasks. Additionally, task-specific training and optimization were conducted for targeted tasks (e.g., standing detection and sitting action detection), further improving the system's application value.

References

[1] Toshev A, Szegedy C. DeepPose: Human Pose Estimation via Deep Neural Networks[J]. CoRR, 2014: 1653-1660.

[2] Kreiss S, Bertoni L , AlahiA. PifPaf: Composite Fields for Human Pose Estimation[J]. CoRR, 2019: 1903-1906.

[3] Vasileios B, Sikandar A, Mykhaylo A, et al. 3D Pictorial Structures Revisited: Multiple Human Pose Estimation.[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 1929-1942.

[4] Liu Yu. Research on Human Posture Recognition Method Based on Improved Deep Learning[J]. Computing Technology and Automation, 2024, 43(02): 182-186.

[5] Zeng Wenxian, Li Yuesong. Deep Learning Algorithm for Key Point Detection of Human Pose Image[J]. Computer Simulation, 2024, 41(05): 209-213,219.

[6] Shan Wei. Human Posture Recognition Algorithm Based on Deep Learning[J]. Heilongjiang Science, 2024, 15(10): 91-93.

[7] Guo Zhenhua, Xie Xuehao, Wang Senlong, et al. Research and Implementation of Tea Bud Recognition Technology Based on Improved YOLOv8[J]. Computer Knowledge and Technology ,2024, 20(14): 23-25.

[8] Ma Le. Recognition and Pose Estimation of Moving Human Body Based on Deep Learning[D]. Chinese Academy of Sciences(Institute of Automation, Chinese Academy of Sciences), 2020.

[9] He Peng.3D Human Pose Recognition via Deep Learning[D]. Shijiazhuang Tiedao University,2021.

[10] Lian Jingxiang. Research on Human Pose Estimation Based on Deep Learning[D]. South China University of Technology, 2021.

[11] Yang Haihong. Deep Learning Based Pose-driven Feature Learning Method for Video Human Action Recognition]. Journal of jiujiang University (natural science edition), 2022, 37(02): 59-64.

[12] Yang Guangyao. Design and Implementation of a Deep Learning-Based Human Pose Recognition System[J]. China New Technologies and Products Journal, 2022, (07):22-24.