WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n1.23
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Pose Estimation Algorithm for Construction Workers

Zhihao Ni [1, 2], Hong Song [3], Hao Wu [1, 2]

[1] School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644000, China

[2] Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science and Engineering, Yibin 644000, China

[3] Aba Teachers University, Aba, Sichuan 623002, China

ABSTRACT

In order to solve the problem of difficulty in judging the human posture of construction workers in complex environments, this paper uses the method based on bone points to identify the human posture, first produces a video dataset of construction workers, and then compares the two algorithms, and selects a suitable pose estimation algorithm for construction site scenarios, which contributes to the pose estimation of construction workers.

KEYWORDS

Construction; DARK_HRNet; Pose estimation.

1. INTRODUCTION

Multi-person pose estimation can be divided into two mainstream methods: top-down and bottom-up [1]. The top-down method first detects all the targets from the input image, captures the target position, and then uses the pose estimation algorithm to detect all the bone points of the intercepted target. DeepPose [2] uses a cascade regressor to refine the prediction results, and applies deep learning to the field of pose estimation for the first time. Due to the lack of AlexNet's feature extraction ability, Chen [3] used the GlobalNet network as the backbone network and changed the model to a two-stage detection in the Cascaded Pyramid Network (CPN), first using GlobalNet to locate simple bone points, and then using RefineNet [4] to integrate the multi-level features of GlobalNet to detect bone points that are difficult to identify and occlude. AlphaPose [5] designed the Symmetric Spatial Transformer Network (SSTN) to enable the detected object to be placed in the center of the detection frame to obtain a high-quality single area, then used SPPE to output the human suggestion box, and finally used P-NMS (Parametric Pose No-Maximum-Suppression) to eliminate redundant poses and make the recognition more accurate. The bottom-up method directly identifies the bone points in the image, and then groups the bone points through the algorithm. The downside of the bottom-up approach is that it doesn't take advantage of global contextual information. OpenPose [6] focuses on the vectors of limb position and direction and the confidence of bone points through Part Affinity Fields and Part Detection Confidence Maps, respectively, and then uses these two branches to learn the connection between bone points, and finally uses Greedy parsing Algorithm to encode the global context, so that the algorithm can basically achieve real-time detection.

2. HUMAN POSE ESTIMATION ALGORITHM

2.1. Introduction to Human Joints

In order to make the pose estimation algorithm accurately detect human bone points, this paper uses the COCO dataset to train the network, and constructs the construction worker behavior dataset through the joint point data trained by the human pose estimation network. The COCO dataset uses 17 key points to estimate human posture, as shown in Figure 1, which are: Nose, Left eye, Right eye, Left ear, Right ear, Left shoulder, Right shoulder, Left elbow, Right elbow, Left wrist, Right wrist, Left hip, Right hip, Left knee, Right knee, Left ankLe, Right ankle.

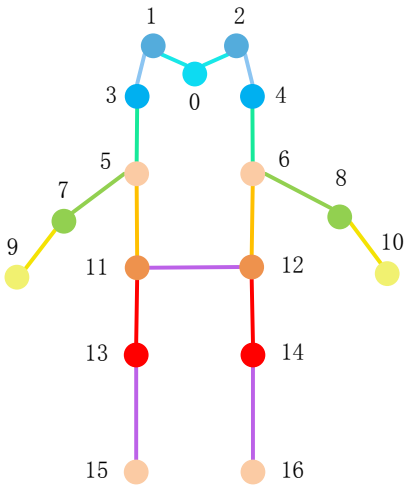

Figure1. Human joint points and joint point associations

3. THE NETWORK OF HUMAN POSE ESTIMATION

In order to choose an algorithm suitable for the construction site scene to perform the bone point detection task, this paper selects two algorithms to compare the two methods, so as to choose the more suitable pose estimation algorithm.

The algorithm needs to use the object detection algorithm to detect the human body first, and then use the human pose estimation algorithm to estimate the pose after cropping the human target. In this paper, the PP-YOLOE [7] algorithm was used for human object detection, and the DARK_HRNet [8] was used for human pose estimation.

3.1. PP-YOLOE

PP-YOLOE is mainly divided into three parts: backbone network, neck, and detection head. The backbone network is responsible for the initial extraction of features, the neck part fuses the features of different levels, and the detection head part integrates the information.

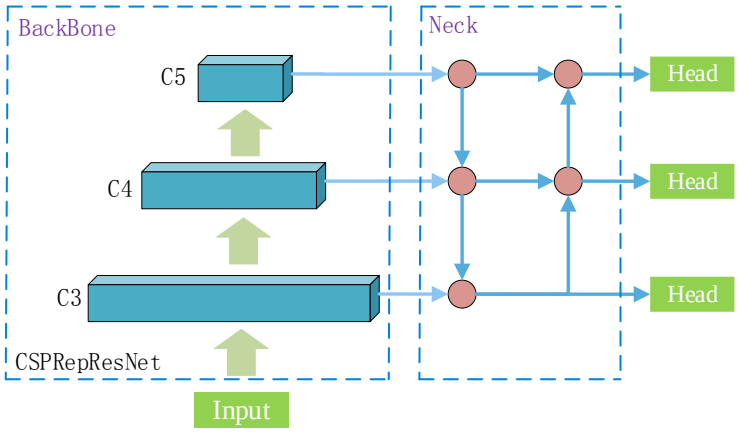

Figure 2. The network architecture of PP-YOLOE

3.2. DARK_HRNet

DARK_HRNet is a high-precision human pose estimation model, which uses new coding and decoding strategies to improve the detection accuracy of human bone points, and the network can be divided into three parts: parallel convolutional flow, feature fusion, and detection head.

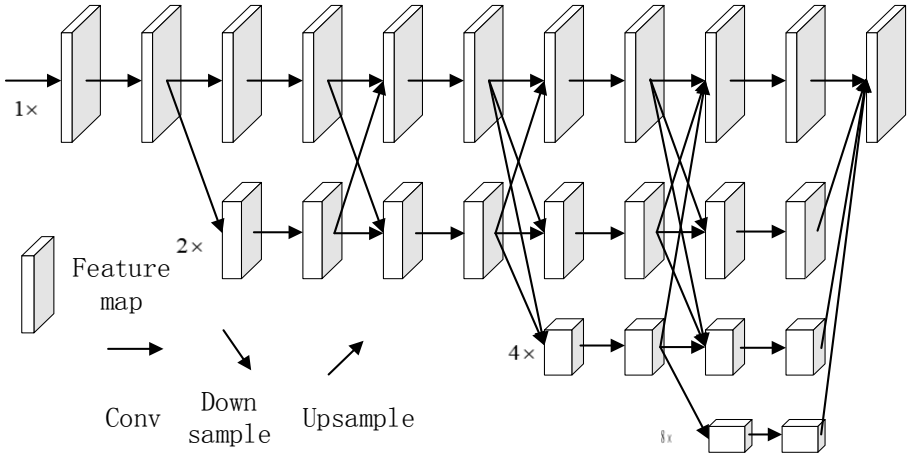

Figure 3. The network architecture of DARK_HRNet

3.3. HRNet_DEKR

The backbone network of the HRNet_DEKR follows the structure of HRNet, and its main contribution is to design an adaptive convolution, which uses affine transformation to recalculate the convolution kernel of the convolution, the ordinary convolution can only learn the center pixel and less surrounding pixels, and the new convolution kernel can expand the range of learning pixels and concentrate on learning the pixels around the key point area.

In order to improve the detection accurac of key points, DEKR uses a separate regression strategy to output the feature map of the backbone network as K feature maps, and divide them into K branches according to the feature map, each branch can learn its own adaptive convolution. Multi-branch learning can decouple features between different key points, thereby improving the quality of regression.

4. COMPARISON OF POSE ESTIMATION MODELS

4.1. Experimental Environment

The hardware configuration of this experiment is as follows: CPU (Intel(R) Xeon(R) Silver 4210), GPU (NVIDIA GeForce RTX 2080 Ti), software environment is: python 3.7, paddle 2.6.0, and the weight is the pre-training weight of COCO dataset.

4.2. Video Dataset Visualization

Figure 4 shows the partial application of the algorithm to the self-made dataset, and it can be seen that the four images of the HRNet_DEKR on the left are compared with the DARK_HRNet on the right, and the bone points are missing to a certain extent. The human body in Figure (a) is a back-to-back perspective of climbing over the fence, and the bone points in the left ear and left eye of the human body are occluded. Figure (b) shows the back-body perspective, and HRNet_DEKR the bone points in the right arm area, the right ear and the nose of the human body are missed, and there is no missing detection in the DARK_HRNet algorithm. Figure (c) is a frontal view, because the relative image size of the left ear is too small, HRNet_DEKR cannot return to the bone point, so there is no missed detection, DARK_HRNet there is no missed detection. Figure (d) shows that the main face area and the left leg area are occluded, and the HRNet_DEKR algorithm is missing.

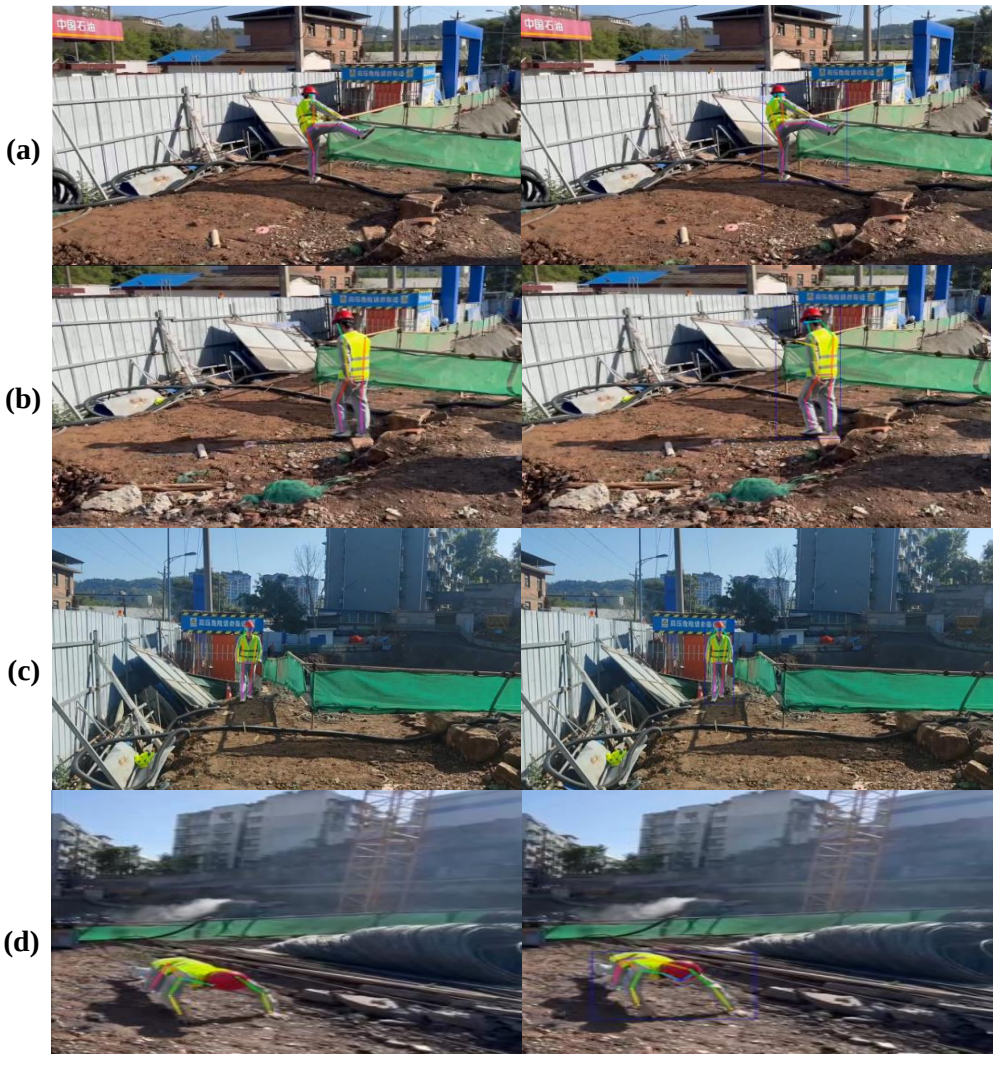

Figure 4. Self-made dataset pose estimation results

Figure 5 shows the comparison of the pose estimation effect using multi-person scenes, in Figure (a), the HRNet_DEKR algorithm mainly misses the bone points in the head area and arm area of the three human targets, and the HRNet_DEKR is that the bone points are identified globally in the picture, so there is a large amount of interference information, and it is impossible to correctly estimate and identify the difficult bone point targets. DARK_HRNet only a part of the human frame in the image is predicted, it is less difficult to estimate and identify bone points based on the characteristics of other parts of the human body, so human bone points can be detected accurately. DARK_HRNet algorithms can detect correctly. In Figure (b), the HRNet_DEKR algorithm mainly focuses on the missing detection of the bone points in the facial area of the person on the left and the bone points of the left hip and left ear of the person on the right. In Figure (c), the HRNet_DEKR algorithm misses a human target, and there is a bone point missing in the left half of the human body. In Figure (d), the HRNet_DEKR algorithm misses the detection of the right human target and the bone points of the left human lower limb and head area, and the left arm area of the middle human body also misses the detection, and the DARK_HRNet algorithm does not miss the detection.

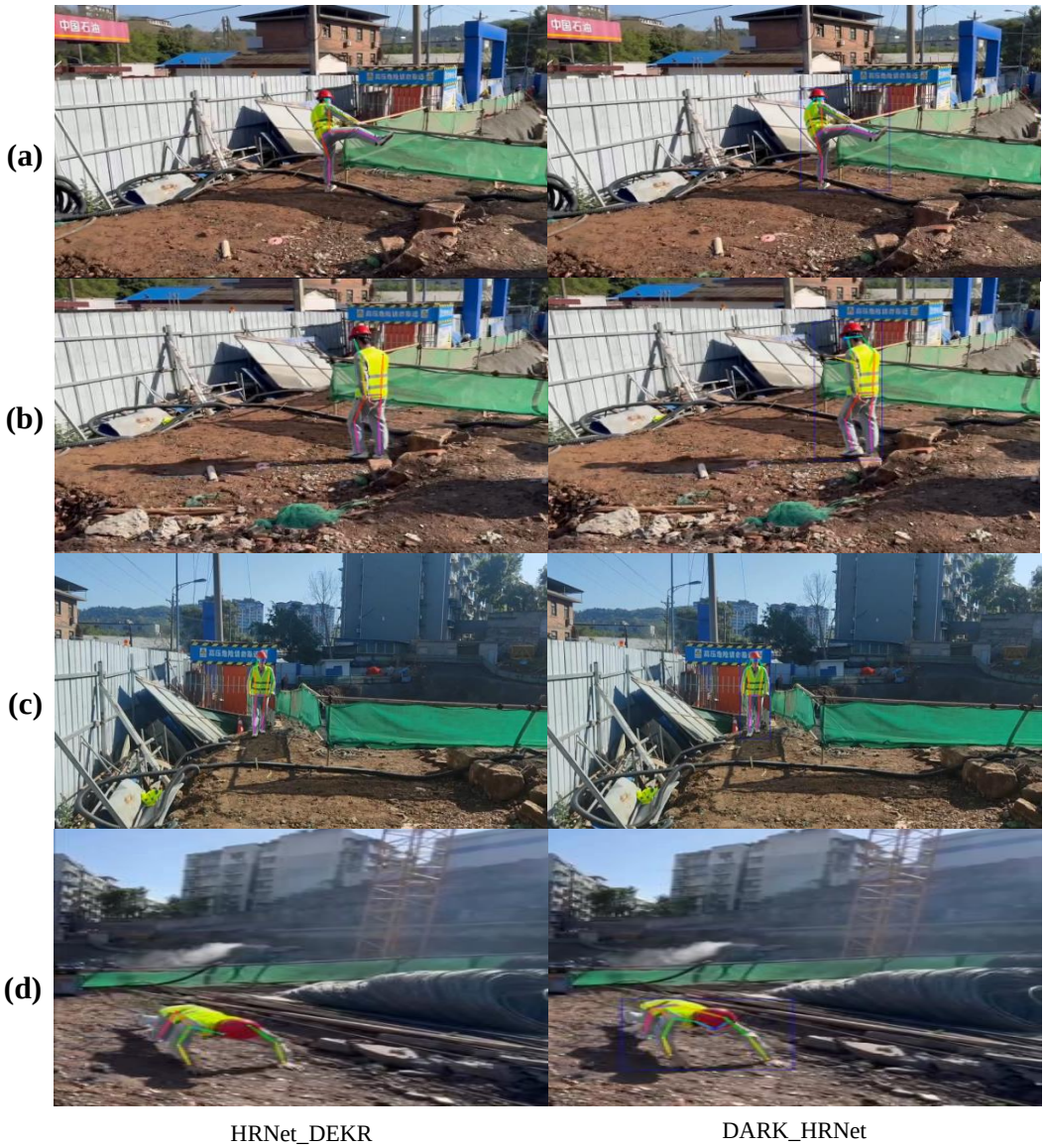

Figure 5. Self-made dataset pose estimation results

In summary, the HRNet_DEKR is very serious for the detection of occlusion and small bone points, which indicates that the robustness of the algorithm is insufficient, and there are great limitations in the detection of incomplete features or serious interference. DARK_HRNet can infer and fill the

features of the complete target based on part of the information, and can maintain high accuracy of object detection and bone point recognition.

5. CONCLUSION

By comparing the two pose estimation algorithms, this paper finds that PPYOLOE and DARK_HRNet can better identify the bone point information of the human body in the construction worker scene, which has a certain application value in the construction site scene, HRNet_DEKR there are many missed detections, which are not suitable for the construction site scene, and the detection accuracy needs to be further improved.

ACKNOWLEDGEMENTS

Sichuan University of Light Industry Graduate Student Innovation Fund (Y2022136).

REFERENCES

[1] Xiaohe Li, A Review of Two-Dimensional Human Pose Estimation [J]. Modern Computer, 2019, No.658(22):33-37.

[2] Toshev A, Szegedy C. Deeppose: Human pose estimation via deep neuralnetworks [C]//Proceedings of the IEEE conference on computer vision and patternrecognition. 2014:1653-1660.

[3] Chen Y, Wang Z, Peng Y, et al. Cascaded pyramid network for multi-person poseestimation [C]//Proceedings of the IEEE conference on computer vision and patternrecognition.2018:7103-7112.

[4] Lin G, Milan A, Shen C, Reid I. Refinenet: multi-path refinc-ment networks for high-rcsolution semantic segmcntation. In: Proceedings of the 30th IEEE Conference on Computer Visionand Pattern Recognition.Honolulu, USA:IEEE, 2017.5168-5177

[5] Fang H S, Xie S, Tai Y W, et al.Rmpe. Regional multi-person pose estimation [C]/IEEE InternationalConference on Computer Vision. Piscataway: IEEE Computer Society, 2017:2334-2343.

[6] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7291-7299.

[7] Xu, S., Wang, X., Lv, W., Chang, Q., Cui, C., Deng, K., ... & Lai, B. (2022). PP-YOLOE: An evolved version of YOLO. arXiv preprint arXiv:2203.16250.

[8] Zhang F, Zhu X, Dai H, et al. Distribution-aware coordinate representation for human pose estimation [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 7093-7102.