2024, 21(2 Special Issue): 0755-0763
https://doi.org/10.21123/bsj.2024.9775
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Human Pose Estimation Algorithm Using Optimized Symmetric Spatial Transformation Network

***Shengqing Lin**[1,2], **Nor Azizah Ali**[1], **Azlan Mohd Zain**[1], **Muhalim Mohamed Amin**[1]*

[1]Faculty of Computing, Universiti Teknologi Malaysia (UTM), Johor Bahru, Johor 81310, Malaysia.
[2]School of Computing and Information Sciences, Fuzhou Institute of Technology, Fuzhou, Fujian, China.
*Corresponding Author.
ICAC2023: The 4th International Conference on Applied Computing 2023.

Received 30/09/2023, Revised 10/02/2024, Accepted 12/02/2024, Published 25/02/2024

Abstract

Human posture estimation is a crucial topic in the computer vision field and has become a hotspot for research in many human behaviors related work. Human pose estimation can be understood as the human key point recognition and connection problem. The paper presents an optimized symmetric spatial transformation network designed to connect with single-person pose estimation network to propose high-quality human target frames from inaccurate human bounding boxes, and introduces parametric pose non-maximal suppression to eliminate redundant pose estimation, and applies an elimination rule to eliminate similar pose to obtain unique human pose estimation results. The exploratory outcomes demonstrate the way that the proposed technique can precisely recognize the human central issues, really work on the exactness of human posture assessment, and can adjust to the intricate scenes with thick individuals and impediment. Finally, the difficulties and possible future trends are described, and the development of the field is presented.

Keywords: Computer vision; deep learning; human post estimation; key point recognition; symmetric spatial transformation.

Introduction

Human pose estimation (HPE) research involves detecting the location of key joints in two or three dimensions as a basis for identifying and reconstructing parts of the human body [1,2]. Benefiting from a great deal of research in artificial intelligence in the last decade, human pose assessment has gradually become a topical problem in the areas of motion judgment, pose capture, and robotic interaction [3,4].

Reliance on manual labeling is the main feature of traditional human pose estimation algorithms, but the accuracy is poor because it is a direct regression to the coordinates [5]. Significant progress in CNN-based pose estimation research has been made after 2013 in response to the development of Convolutional Neural Networks (CNNs) and the publicly available pose datasets and the precision and speed have been significantly enhanced compared to the traditional algorithms [6]. Compared with traditional

2024, 21(2 Special Issue): 0755-0763
https://doi.org/10.21123/bsj.2024.9775
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

human posture estimation methods, CNN-based methods can learn more adequate representations from the data, so how to utilize CNNs to enhance the performance of pose assessment has become a popular research direction [7].

Related Works

HPE is a complicated task that incorporates both two- and three-dimensional estimation algorithms [8]. 3D pose estimation is based on 2D methods with the addition of relative depth data, and the same two-dimensional pose might represent several three-dimensional ones, which is an inherent problem that needs to be solved for estimating 3D pose estimation from images or videos [9].

Posture detection is categorized into two types: single and multiple joint point inspection [10]. Nevertheless, the core of top-down and bottom-up approaches starts with the detection of a single keypoint [11].

Single-Person HPE Methods

Deep learning extracts abstract features of data layer by layer through a multi-layer neural network structure, which is better adapted to different data types and application scenarios [12]. In addition, deep learning can be robust to noise, occlusion, and other disturbing factors, which greatly improves the accuracy of joint detection.

Single-person HPE is used to inspect one person's key point information in the scene using three forms: one is an approach based on coordinate regression; the second based on heat map detection; the last one is a regression and detection hybrid model Fig.1 [13].

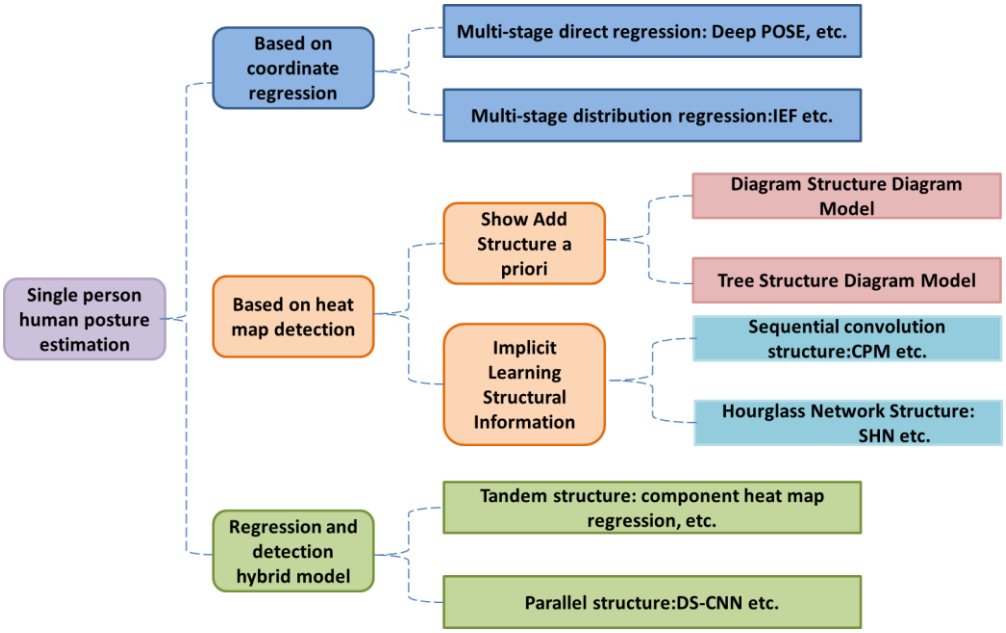

Figure 1. Single human posture estimation method summary

According to the technical approach, pose estimation using convolutional neural networks is categorized into two types: coordinate-based regression and heatmap-based regression [14,15]. Coordinate-based regression methods directly return the coordinates of the key points and compute the probability value corresponding to each pixel in image [16].

2024, 21(2 Special Issue): 0755-0763
https://doi.org/10.21123/bsj.2024.9775
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

In 2014, DeepPose proposed by Toshev et al [17] is the first convolutional neural network for estimating a pose-based model that learns the location of critical points in an image directly without a mannequin or part detector. Predicting key point locations directly from the scene is extremely hard and requires more powerful and robust network models to be introduced. In 2017, Sun et al. introduced a neural network regression method using ResNet-50 as a base, which employs a reparametrized gesture representation that utilizes the joint-connected structures to define a component loss function for encoding remote interaction information [18]. Using only sparse keypoint information is lacking in robustness, and converting heatmap supervision to keypoint supervision retains the benefits of both methods. Luvizon et al. proposed a pose regression method using Soft-argmax functions to transform heat maps into critical point location information in a differentiable manner, resulting in an end-to-end training framework [19].

In single-person pose estimation, heat map-based regression uses heat maps to represent the location of joint points and then trains a neural network to obtain the detection probability of each pixel point. This method has higher accuracy compared to the traditional manual feature method, so the heat maps regression-based method is more popular.

In 2016, Wei et al. proposed the serialized Convolutional Pose Machine (CPM), which uses multiple VGGNet-based sub-networks to form a cascade network to augment the perceptual scope across the network and model the dependencies between key points. The CPM uses the feature maps and confidence maps generated in the prior period as inputs for the next period. The method can learn rich implicit spatial models, employs multi-stage outputs, and uses intermediate supervision to resolve the gradient vanishing problem resulting from the depth of the network.

Multi-Person HPE methods

Unlike single-person HPE methods, multi-person methods face greater challenges as it requires not only determining how many people are in the same scene and where each person is located but also grouping joints belonging to different human bodies. Depending on the solution, multi-person methods are categorized as top-down and bottom-up.

Top-Down Evaluation Method

This kind of methodology recognizes each individual from the image background by a target detection algorithm and then performs skeletal joint point detection, which consists of two processes: joint point detection and joint point clustering. Because of the characteristic of the top-down approach, it is more likely to have duplicate detection and false estimation when the human bodies are occluded from each other in the image.

Sun et al. introduced a High-Resolution Network (HRNet) and iteratively fused features produced by multi-scale sub-networks to produce robust high-resolution [21]. For more precise key point positioning, Cai et al [22] introduced a method to effectively fuse characteristics with the same size to obtain fine localized features, which retain rich low-level spatial information conducive to the precise localization of key points.

Bottom-Up Evaluation Methods

This kind of methodology has two constituent processes: key point detection and key point clustering, i.e. It first detects the entire skeletal joints followed by clustering all these joints as corresponding individuals. The DeepCut network proposed by Pishchulin et al is the first HPE method using the bottom-up method [23]. The method first locates all human key points labels each keypoint, and then solves the key point association problem using the Integer Linear Program (ILP) algorithm. Cao et al. Introduced OpenPose which is a more efficient multi-person approach using heat maps to predict posture coordinates and associates key points with each human body via Part Affinity Field (PAF) [24].

Newell et al. presented the Stacked Hourglass Network (SHN) [25]. This algorithm allows the hourglass modules to learn information from the previous hourglass module and refine body-related scaling features, allowing each module to generate a

2024, 21(2 Special Issue): 0755-0763
https://doi.org/10.21123/bsj.2024.9775
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

complete heat map, thereby improving joint prediction accuracy.

As a target detection method for CNN networks, Faster R-CNN [26] integrates characteristic extraction, suggestion extraction, box selection regression (rectangular refinement), and classification, significantly improving the performance.

Transformers [27] are currently extremely popular, and a large number of Transformer-based algorithms, such as ViTPose [28], have been generated in HPE, which can collect global information and improve accuracy. However, the disadvantage of these algorithms is that they must perform massive computations on huge human posture datasets, which is highly demanding in terms of computational power. They are not suitable for some of the speed-demanding scenarios, such as 2D video surveillance, motion detection, etc. How to optimize the CNNs algorithm to extract the human body frames quickly and reduce arithmetic power usage is the goal of this research.

Methods

In this study, an optimized symmetric spatial transform network for 2D pose estimation is proposed, which can eliminate similar poses generated by CNNs, and a high-quality human target frame is proposed to improve the bounding box accuracy. It can handle different poses and orientations, ensuring that the network can adaptively adjust the bounding box to a wide range of human poses and orientations. Based on the simplified version of the convolutional neural network (Fast R-CNN) architecture, the entire symmetric space and transformation network framework is designed, using SSD [29] (Single Shot Multibox Detector) +SHN detection.

The first step of this method introduces parametric pose non-maximum suppression to eliminate redundant pose estimation, the second step uses SHN to fit the single-person pose for estimation, and the third step applies elimination of similar poses to obtain unique human pose estimation results, to get unique human body position estimation results, and finally SSD is used as a target detection algorithm for detecting the human body target.

The model takes images of arbitrary size as input and uses an attention-based module to pair find possible points of interest, performs multilayer convolutional computation, generates a sliding window on the image to be convolved, and then inputs a fully connected layer for regression. The number of convolutional layers that can be shared in this module is set to seven to balance accuracy and efficiency.

N candidate boxes are extracted from each inference score map S_j. However, when estimating each candidate box, incorrect estimation will occur due to blur or similar background color. Therefore, integer linear programming was introduced to connect each joint to the correct human body and eliminate erroneous contours. The pose of an individual in a given image is defined as $\chi=\{\boldsymbol{X}_j\}_{j=1,\ldots,J}$, where J denotes 1-14 junction. The j_{th} junction's location within a scene is represented as $\boldsymbol{X}_j \in x$.

The background information from the previous stage is utilized in each following stage and generates a new confidence score:

$$\varphi_t > 1[\boldsymbol{X}|\boldsymbol{I}, \Psi(\boldsymbol{X}, \boldsymbol{S}_{t-1})] \rightarrow \{s_t^j(\boldsymbol{X}_j = \boldsymbol{X})\}_{j=1,\cdots,J+1} \quad (1)$$

where: $\boldsymbol{S}_t \in R^{w \times h \times (J+1)}$ denotes the set of confidence scores for all joints in stage t and $\psi(\boldsymbol{X}, \boldsymbol{S}_{t-1})$ denotes the feature mapping starting from the confidence mapping $\boldsymbol{S}_{t-1}$ to position $\boldsymbol{X}$. Therefore, instead of taking the maximum value of the confidence score, *N* candidates are drawn from each inferred score map $\boldsymbol{S}_j$.

The second step is separated into two phases: first phase, multiple bodies are detected using Faster-RCNN, and each body bounding box is cropped while maintaining the human body aspect ratio of 1.37, resulting in a cropped image size of 353 × 257. In the second stage, a fully convolutional network is used to predict the heat map and offset of the body in each body bounding box, and the precise location of the joint points is obtained by fusing the heat

2024, 21(2 Special Issue): 0755-0763
https://doi.org/10.21123/bsj.2024.9775
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

map and offset map. The offset map represents the offset from each pixel position of the heat map to the correct node position (each offset map correlates into one channel, denoting the x-coordinate and y-coordinate). The three channels are fused to vote the true node position from the predicted offset. The fusion is performed as follows.

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_c}\sum_i L_c(p_i, p_{i^*}) + \alpha\frac{1}{N_r}\sum_i p_{i^*} L_r(t_i, t_{i^*}) + \beta \quad 2$$

where i is the index of the anchor of each bounding box, and p_i is the predicted probability of the box with index i. The predicted probability of the box with index i is the predicted probability of an object, p_{i^*} is the probability that the object is correct, t_i is a vector representing the parameterized coordinates of the predicted bounding box, and t_{i^*} is the coordinates of the actual box associated with the positive anchor. The term α is a balancing parameter that weighted the result L. The term β is a bias, optional hyper parameter. L_c is the logarithmic loss. L_r was used to represent the regression loss.

$$f_k(x_k) = \sum_j \frac{1}{\pi R^2} G\left[x_j + F_k(x_j) - x_i\right] h_k(x_j) \quad 3$$

Experiments and Analysis

Data set creation

In this study, a public and highly mainstream dataset for evaluating human pose recognition, MPII Human Pose, is used to test and refine the system. This dataset contains images obtained from a vast number of human action videos with joint nodes and action annotations. 10,000 images from MPII Human Pose are selected for network training, and 1000 images are selected for testing the trained network model.

In addition, data expansion is performed on selected datasets to enhance the generalization ability of the network model. In this paper, random horizontal and vertical flipping with a range of ±45° of random swivel and random scaling to [0.7, 1.35] were used for filtering the images for training the network model. By the above data enhancement methods, the complexity of the samples in the data set is increased and overfitting of the model can also be avoided.

Experimental platform and performance evaluation index

The experimental environment is set up on an Ubuntu 16.04 system with Inter(R) XeonSilver4110 CPU, NVIDIA GeForce RTX 3080Ti GPU, and Pytorch deep learning framework.

The object keypoint Similarity (*OKS*) has been chosen for measuring comparability between the expected keypoints with the real keypoints, which is calculated by the following formula:

$$R_i = \sum_i exp\{-d_{pi}^2/2S_p^2\sigma_i^2\}\delta(v_{pi}=1) \quad 4$$

$$OKS_p = \frac{R_i}{\sum\delta(v_{pi}=1)} \quad 5$$

$$OKS_p = \frac{\sum_i exp\{-d_{pi}^2/2S_p^2\sigma_i^2\}\delta(v_{pi}=1)}{\sum\delta(v_{pi}=1)} \quad 6$$

Where: p denotes the ID of the human body; i denotes the ID of the key point, d_{pi} represents Euclidean distance among predicted ith critical spot of the pth human body and authentic labeled key point; S_p denotes the area occupied by the target boundary of the pth body; σ_i is the key point normalization factor, which is a constant and denotes whether the ith critical spot of ith human body is visible or not; and δ is the selection function.

Training and Testing

During training, an error update algorithm was chosen from Adam, and the initial study percentage was set to 0.0001, the study percentage was reduced to 0.5 with 20 iterations, and set batch length to 8 with smaller data, for 80 rounds of training iterations.

In the testing phase, the Stacked Hourglass Network **(**SHN) HPE network is selected for the single-person network and the SSD algorithm is selected to

2024, 21(2 Special Issue): 0755-0763
https://doi.org/10.21123/bsj.2024.9775
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

detect human targets. To ensure that the target frame obtained by object detention algorithm can cover entire body area, the detected human target frame is extended by 15% in both height and width directions.

Results and Discussion

On the MPII [30] data set, the precision of the strategy worked on by 0.8% over DeepCut, and the computation time improved from 57995s to 10s. To assess the impact of the recognition precision of the detector on the HPE, the accuracy was significantly improved from 49.3% to 76.9% when the true position of the detected human body was given, showing that a superior human identifier would additionally further develop the human posture assessment results.

Another set of human detection algorithms and a single HPE network are included to demonstrate the robustness of the calculation presented in this paper.

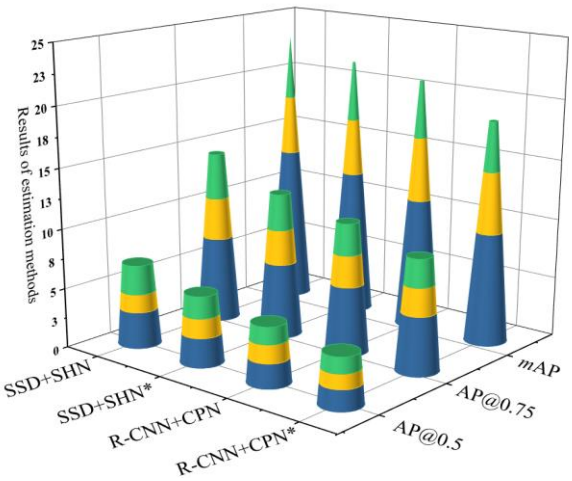

Figure 2. Presentation of the proposed calculation

Faster R-CNN and the CPN pose estimation algorithm are chosen to compare the presentation of the first calculation and the proposed calculation after adding this paper, respectively, and the results are shown in Fig.2, with "*" denoting the addition of the algorithmic framework presented in this paper. AP means Average Precision and mAP means mean Average Precision (Average of precision for all categories). AP@0.5 (0.75) represents the average prediction with the IOU (Intersection-Over-Union) greater than 0.5(0.75). AP in the pose estimation is defined by the OKS. For example, AP0.5 (AP at OKS = 0.50). From the results of the tests in Fig 2, it tends to be seen that in the wake of adding the calculation proposed, two sets of pose estimation algorithms mAP improved the average accuracy by 4.5% and 3.4%, which shows the viability of the proposed calculation and can successfully work on the exactness of the hierarchical posture assessment calculation.

Ablation Experiment

To investigate effectiveness of each proposed symmetric spatial transformation network, parallel single-person pose estimation network, and pose non-maximal suppression, the SSD target detection algorithm and SHN single-person pose estimation algorithm are used as examples, and the following experimental schemes are designed to test each group of experimental schemes separately, and the exploratory outcomes are displayed in Fig.3.

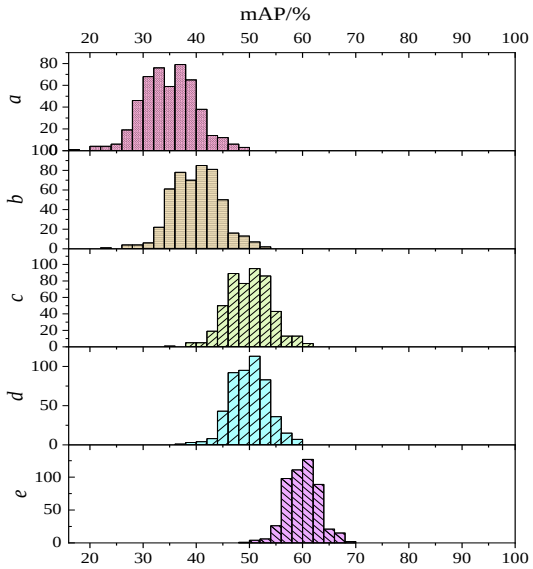

Figure 3 Relationship of the experimental schemes designed to test each group of experimental schemes

Comparing experimental scenarios *a*, *c*, and *e*, it very well may be seen that the exactness of the calculation diminishes by 2.2% when the symmetric spatial change organization and the equal single-individual posture assessment network are missing,

2024, 21(2 Special Issue): 0755-0763
https://doi.org/10.21123/bsj.2024.9775
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

demonstrating the adequacy of the proposed symmetric spatial change organization and the equal single-individual posture assessment organization, i.e., a high-quality human target frame can be generated for subsequent human pose estimation; comparing experimental scenarios *a*, *b*, and *e*, the accuracy of the algorithm decreases by 0.6% when the single-person HPE is missing, indicating the effectiveness of the proposed single-person pose when comparing experimental schemes *a*, *b*, and *e*, the accuracy of the algorithm decreases by 0.6% when the parallel single-person HPE is missing, indicating the effectiveness of the proposed single-person pose, i.e., it can more readily move the human body to the focal point of the objective proposed locale for exact posture assessment; when comparing experimental schemes *a*, *d*, and *e*, the accuracy of the algorithm decreases by 0.7% when the pose non-maximal suppression is missing, indicating the viability of the proposed pose non-maximal suppression, i.e., the redundant poses can be effectively eliminated.

A multi-person pose estimation algorithm is proposed to join the symmetric spatial change network with an equal single-individual posture assessment organization to propose high-quality human target frames from inaccurate human bounding boxes because the presentation of the human objective discovery algorithm easily affects the accuracy of human key point localization. The experiment results demonstrate that the approach suggested in this research does, in fact, improve the accuracy of HPE.

Conclusion

Despite the fact that pose estimate studies have made substantial advances in recent years attributed to the persistent work of numerous researchers, the existing algorithms still present significant challenges. This paper focuses on how to optimize the algorithm in the pose estimation task for solving the invalid problem of key point localization accurately affected by the performance of detection algorithms, and how to increase the precision and accuracy in crowded situations. An optimized symmetric spatial transformation network algorithm is proposed for the case of inaccurate and redundant detection of human target detection algorithms, which integrates a symmetric spatial transformation model to a concurrent pose assessment network to propose a first-class human target shell from an inaccurate human bounding box and introduces parametric pose non-extreme value suppression eliminates the redundant pose estimation. The results of the experiments indicate that the methods presented in this paper effectively enhance accuracy of human body pose estimation. In the future, HPE will occupy a more important position in the future real life as an important part of computer vision.

Acknowledgment

We would like to thank the professors and doctors from the Universiti Teknologi Malaysia who gave their feedback on the paper.

Authors' Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at University Teknologi Malaysia.

Authors' Contribution Statement

L. S., N. A. A., A. M. Z. and M. M. A. designed the study. L. S. performed the experiments and analyzed the data. L. S., N. A. A., A. M. Z. and M. M. A. wrote the paper with input from all authors.

2024, 21(2 Special Issue): 0755-0763
https://doi.org/10.21123/bsj.2024.9775
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

References

1. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, et al. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing. 2020; 408(2): 189-215. https://doi.org/10.1016/j.neucom.2019.10.118.
2. Alzubaidi L, Zhang J, Humaidi A J, et al. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. J. Big Data. 2021; 8(1): 1-74. https://doi.org/10.1186/s40537-021-00444-8.
3. Pareek P, Thakkar A. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. Artif. Intell. Rev. 2021; 54(3): 2259-2322. https://doi.org/10.1007/s10462-020-09904-8.
4. Shi Y, Zhang Z, Huang K, et al. Human-computer interaction based on face feature localization. J. Vis. Commun. 2020; 70(1): 102740. https://doi.org/10.1016/j.jvcir.2019. 102740.
5. Zheng C, Wu W, Yang T, et al. Deep learning-based human pose estimation: A survey. Published online arXiv:2020;13392.https://doi.org/10.48550/arXiv.2012.13392.
6. Chen J, Li S, Liu D, et al. Indoor camera pose estimation via style-transfer 3D models. COMPUT-AIDED CIV INF. 2022; 37(3): 335-353. https://doi.org/10.1111/mice.12714.
7. Li M, Gao Y, Sang N. Exploiting learnable joint groups for hand pose estimation. Proceedings of the AAAI Conference on Artificial Intelligence. 2021; 35(3): 1921-1929. https://doi.org/10.1609/aaai.v35i3.16287.
8. Tang H, Wang Q, Chen H. Research on 3D human pose estimation using RGBD camera. 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC). 2019: 538-541.https://doi.org/10.1109/iceiec.2019.8784591.
9. Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale aware representation learning for bottom-up human pose estimation. arXiv. 2020. https://doi.org/10.48550/arXiv.1908.10357.
10. Jin S, Liu W, Xie E, et al. Differentiable hierarchical graph grouping for multi-person pose estimation. European Conference on Computer Vision. 2020; 718-734. https://doi.org/10.48550/arXiv.2007.11864.
11. Bao Q, Liu W, Cheng Y, et al. Pose-guided tracking-by-detection: Robust multi-person pose tracking. IEEE Transactions on Multimedia. 2020; 23(10): 161-175.https://doi.org/10.1109/TMM.2020.2980194.
12. Dang Q, Yin J, Wang B, et al. Deep learning based 2d human pose estimation: A survey. Tsinghua Sci Technol. 2019; 663-676. https://doi.org/10.26599/TST.2018.9010100.
13. Luvizon D C, Picard D, Tabia H. 2d/3d pose estimation and action recognition using multitask deep learning. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5137-5146.https://doi.org/10.1109/CVPR.2018.00539.
14. Chen Y, Tian Y, He M. Monocular human pose estimation: A survey of deep learning-based methods. Comput Vis Image Underst. 2020; 192(5): 102897. https://doi.org/10.1016/j.cviu.2019.102897.
15. Liu, X., Zhang, T. and Liu, M. Joint estimation of pose, depth, and optical flow with a competition–cooperation transformer network. Neural Networks. 2024; 263–275. https://doi.org/10.1016/j.neunet.2023. 12.020.
16. Qiu S, Zhao H, Jiang N, et al. Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. Information Fusion. 2022; 80(6): 241-26. https://doi.org/10.1016/j.inffus.2021.11.006.
17. Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEEPress. 2014. 1653-1660. https://doi.org/10.1109/ CVPR.2014.214.
18. Li S, Zhang L, Diao X. Deep-learning-based human intention prediction using RGB images and optical flow. J Intell Robot Syst. 2020; 95-107. https://doi.org/ 10.1007/s10846-019-01049-3.
19. Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation.2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018; 7103–112. https://doi.org/ 10.1109/CVPR. 2018.00742.
20. Wei SE, Ramakrishna V, Kanade T, Sheikh Y. Convolutional pose machines. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; 4724-4732. https://doi.org/2433-2454. 10.1109/CVPR.2016.511.
21. Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019; 5693-5703. https://doi.org/10.1109/CVPR.2019.00584.

2024, 21(2 Special Issue): 0755-0763
https://doi.org/10.21123/bsj.2024.9775
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

22. Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, et al. Learning delicate Local Representations for Multi-person Pose Estimation. In European Conference on Computer Vision (ECCV). 2020; 457-472. https://doi.org/ 10.1109/CVPR.2019.00584.
23. M. Rajchl et al. DeepCut: Object Segmentation From Bounding Box Annotations Using Convolutional Neural Networks. IEEE Transactions on Medical Imaging. 2017; 36(2).674-683. https://doi.org/10.1109/TMI.2016.2621185.
24. Cao Z,Simon T,WeiS H, et al. Real time multiperson 2D pose estimation using part affinity fields. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE Press. 2017; 1302-1310. https://doi.org/10.1109/TPAMI.2020.2983686.
25. Newell A, Yang KY, Deng J. Stacked hourglass networks for human pose estimation. Computer Vision - ECCV 2016. Lecture Notes in Computer Science. 2016; 483-499. Available from: https://doi.org/10. 1007/978-3-319-46484-8_29.
26. Miller LE, Fabio C, Azaroual M, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017; 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031.
27. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017; 5998-6008. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
28. Yufei Xu, Jing Zhang, Qiming Zhang, Dacheng Tao. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. 2022; 38571-38584. https://doi.org/10.48550/arXiv.2212.04246.
29. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: Single shot multibox detector. Computer Vision ECCV(Springer). 2016; 21-37. https://doi.org/10.1007/978-3-319-46448-2_0.
30. Andriluka M, Pishchulin L, Gehler P, Schiele B. 2D human pose estimation: New benchmark and state of the art analysis. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014; 3985-3978. https://doi.org/10.1109/CVPR.2014.471.

خوارزمية تقدير وضعية الإنسان باستخدام التناظر الأمثل شبكة التحول المكاني

شينغكينغ لين[1,2]، نور عزيزة علي[1]، أزلان محمد زين[1]، محلم محمد أمين[1]

[1]كلية الحاسبات، الجامعة التكنولوجية الماليزية(UTM) ، جوهور باهرو، جوهور 81310، ماليزيا.
[2]كلية علوم الحاسب والمعلومات، معهد فوتشو للتكنولوجيا، فوتشو، فوجيان، الصين.

الخلاصة

يعد تقدير وضعية الإنسان موضوعًا بالغ الأهمية في مجال رؤية الكمبيوتر، وقد أصبح نقطة ساخنة للبحث في العديد من الأعمال المتعلقة بالسلوكيات البشرية. يمكن فهم تقدير وضع الإنسان على أنه مشكلة التعرف على النقاط الرئيسية للإنسان والاتصال بها. تقدم هذه الورقة شبكة تحويل مكاني متماثلة محسنة مصممة للتواصل مع شبكة تقدير وضعية الشخص الواحد لاقتراح إطارات مستهدفة بشرية عالية الجودة من الصناديق المحيطة البشرية غير الدقيقة، وتقدم قمعًا بارامتريًا غير أقصى للقضاء على تقدير الوضعية الزائدة عن الحاجة، وتطبق قاعدة الإزالة لإزالة الوضع المماثل للحصول على نتائج فريدة لتقدير الوضع البشري. توضح النتائج الاستكشافية كيف يمكن للتقنية المقترحة أن تتعرف بدقة على القضايا الإنسانية المركزية، وتعمل حقًا على دقة تقييم وضعية الإنسان، ويمكنها التكيف مع المشاهد المعقدة مع الأفراد السميكين والعوائق. وأخيرا، يتم وصف الصعوبات والاتجاهات المستقبلية المحتملة، ويتم عرض تطور المجال.

الكلمات المفتاحية: رؤية الكمبيوتر؛ تعلم عميق؛ تقدير ما بعد الإنسان؛ التعرف على النقاط الرئيسية؛ التحول المكاني المتماثل.

.