DBayC: pose estimation and behavior prediction based on dynamic Bayesian CNN

Rong Zeng, Xiaoshuo Jia*, Qingxuan Lv, Zhihui Li
College of Computer Science, Guangdong University of Science and Technology, Dongguan
523000, Guangdong, China

ABSTRACT

Traditional algorithms have advantages such as interpretability and portability in pose estimation task. However, in complex background environments, traditional algorithms suffer from poor adaptability and detection errors. When dealing with complex scenes or small targets, CNN-based algorithms exhibit superior accuracy compared with traditional algorithms. However, CNN-based algorithms of pose estimation cannot be further developed on mobile terminals due to the large number of model parameters. To address this problem, this paper proposes the DBayC algorithm. First, the LBN (Limb Behavior Network) module is designed based on the CNN (convolutional neural network) algorithm to achieve the semantic segmentation effect on the human body. Then, the node annotation of human body is performed on the semantic segmentation results from LBN module to form graph-structured data. Finally, Bayesian formula is used to perform conditional probability analysis on the nodes in the graph, and the motion trajectories between nodes are analyzed, thereby achieving pose estimation and behavior analysis. Through the training of two data sets Hi-Eve and PoseTrack2017, and comparison with some SOTAs (state of the art) models. The experimental results show that under Hi-Eve data, DBayC achieved an accuracy of 79.2%, which is 3.8% higher than HRNetV2. Under the PoseTrack2017 data set, the DBayC algorithm achieved an accuracy of 78.6%, 6.9% higher than HRNetV2. It can be concluded that not only the accuracy of the DBayC algorithm has been improved, but the portability of the algorithm has also been improved, so the DBayC algorithm has certain use value.

Keywords: CNN, Bayesian, pose estimation, semantic segmentation

1. INTRODUCTION

As the continuous development of artificial intelligence technology, pose estimation and behavior prediction have developed rapidly in various fields. The task of pose estimation and analysis is achieved by marking the joints of the body limbs and linking the limbs. And, the behavioral analysis is achieved through continuous pose estimation. These methods are used in medical diagnosis, human-computer interaction, security monitoring, etc. Although these methods have made some progress, the methods of pose estimation and behavior prediction still suffer some challenges. Pose estimation methods can be divided into single target detection and multi-target detection. Among them, multi-target human pose estimation mainly recognizes the posture of individuals by locating key points of the human body. There are two traditional positioning methods: (1) Top-down positioning: This type of method uses the target detection method to obtain a single instance object firstly, and then extract key point features from the instance. (2) Bottom-up positioning: This type of method first extracts all key feature points from the image without instantiation, and then regresses the extracted feature points to the corresponding instance. Although traditional posture methods (PnP[1], ASM[2], SURF[3]) are simple and interpretable, they exhibit low accuracy and poor adaptability. In deep learning, DeepPose[4,5] was the first to employ convolutional neural networks directly to regress human joint positions from images. OpenPose[6-8] implements multi-person pose estimation based on DeepPose, which can simultaneously detect and locate multiple key points of individuals in images or videos. YoloV7[9] is one of the networks in the Yolo series[10-16]. YoloV7 further implements multi-target pose estimation tasks based on traditional target detection tasks.

Pose estimation algorithms based on deep learning frameworks effectively address issues such as low accuracy and poor adaptability encountered by traditional algorithms. However, its large model, slow speed and poor portability also limit

*gxnujiahsuo@163.com

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024),
edited by Yang Yue, Proc. of SPIE Vol. 13395, 133952U · © 2024 SPIE · 0277-786X · Published
under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3049267

the development of this method. As for those problems, this paper proposes a pose estimation and behavior prediction model based on dynamic Bayesian convolutional neural networks.

This paper first designs the LBN module based on CNN. LBN obtains each instance object from the image and extracts the limb node features Fn about the target from the instance object. Then, an undirected graph is constructed from the target's limb nodes, and the relationship between the nodes in the undirected graph is analyzed using the Bayesian structure. Finally, the behavior is analyzed and identified through the changes in the relationship between the nodes in the undirected graph. DBayC is compared with a series of SOTAs models on two pose detection and behavior prediction datasets. The experiment shows that DBayC can perform pose detection and behavior analysis tasks for multiple targets. The innovation of this paper includes:

(1) This paper introduced Bayesian theory into CNN. Initially, LBN obtains limb node information. Then, the node feature information is used to predict the node information through Bayesian theory. Finally, the predicted node information is regressed to the limb node, thereby achieving the behavior prediction task.

(2) By integrating two modules, this paper not only achieves multi-task processing but also effectively reduces interference from complex noisy environments.

(3) The model algorithm is tested on two data sets. The experiment systematically demonstrates the superiority and generalization of the DBayC mode.

2. DBAYC

In order to accurately realize pose estimation and behavior classification of multiple targets, this paper first uses a deep convolutional network to obtain the node features of the limbs, and then uses a Bayesian network to learn the relevant information between the nodes, and predicts the nodes based on the relevant information, thereby realizing behavior prediction and limb detection.

2.1 LBN

LBN primarily utilizes a CNN structure to achieve semantic segmentation of each instance object, as illustrated in Figure 1. By integrating contextual information, LBN extracts n instance objects, segip {*i*=1, 2 ,3, ..., *n*} from image P.

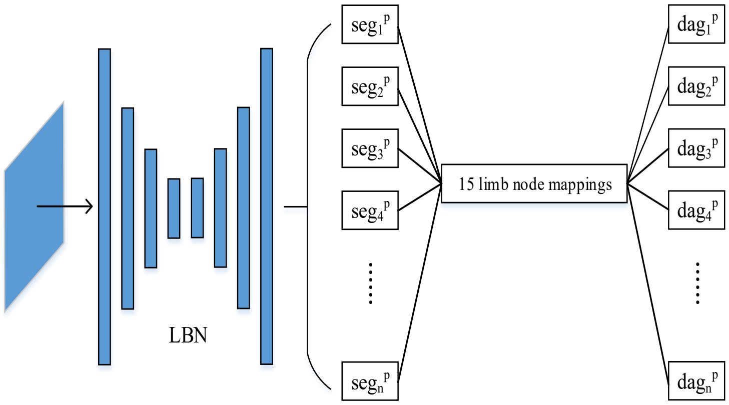

Figure 1. LBN structure.

Then, 15 limb nodes are mapped to each instance object, and the structure of nodes and edges is initialized, thereby obtaining n Directed Acyclic Graphs (DAGs) dagip {*i*=1, 2, 3, ..., *n*} corresponding to *n* instance objects, as depicted in Figure 2.

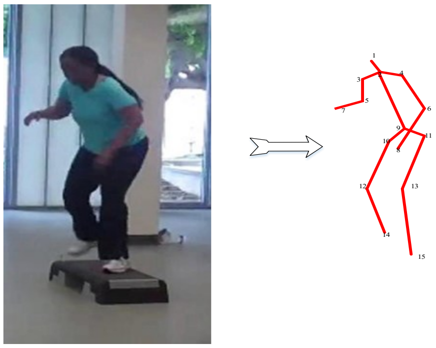

Figure 2. 15 limb nodes mapping.

2.2 Bayesian prediction

The joints and limbs of the human body can also be regarded as a DAG, in which the nodes represent the joints of the body, and the directed edges symbolize the limbs between the joints. The movement of the limb is related to the causal relationship between the two joints. The analysis of DAG leads to the analysis of the patterns of human limb movements and achieve the behavior prediction and analysis task.

Bayesian network is a probabilistic modelling structure that mainly deals with DAG. G(V,L), V is the set of nodes and L is the set of edges. The Bayes formula, as equation (1), describes the probabilities of the distributions of the three nodes a, b, c. And the node determination of G can be divided into 3 cases as show in Figure 3. When it is a precondition of node c that triggers a, b node to change, it can be expressed as equation (2). When a, b, c is the second case of Figure 3, it can be expressed as equation (3). Similarly, when a, b, c structure is the third case of Figure 3, it can be expressed as equation (4).

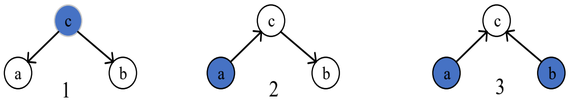

Figure 3. Three cases of node distribution.

$$p(a,b,c)=p(c|a,b)\ast p(b|a)\ast p(a) \tag{1}$$

$$p(a,b,c)=p(c)\ast p(a|c)\ast p(b|c) \tag{2}$$

$$p(a,b,c)=p(a)\ast p(c|a)\ast p(b|c) \tag{3}$$

$$p(a,b,c)=p(a)\ast p(b)\ast p(c|a,b) \tag{4}$$

2.3 DBayC

DBayC algorithm is designed on the basis of LBN module combined with Bayesian theory. The implementation of the algorithm can be divided into the following three steps:

(1) First, the LBN module is used to perform semantic segmentation on the input image P and extract features. Then, the joints and limbs of each character instance object are extracted from the image P.

(2) The joints and limbs extracted above are used as labels, and a DAG of the human body is constructed. The DAG is initialized and assigned values according to the image information of the first frame.

(3) DBayC to predict the behavior through the change of nodes and edges of the DAG.

3. EXPERIMENT

3.1 Experimental environment

DBayC built on the TensorFlow framework. Server configuration with AMD Athlon(tm) II X4 640 Processor x4, NVIDIA GeForce GTX 2070 GPU and Ubuntu 16.04 system. Some SOTA models are selected for comparison here in this paper, such as PoseNet[17], DensePose, MoveNet[18], HRNet[19], HRNetv2[20], and the accuracy is used as the final evaluation criterion. Two datasets, PoseTrack2017[21] and Hi-Eve, were selected here for this paper.

3.2 Experimental comparison of Hi-Eve dataset

The Hi-Eve dataset is a human-centric dataset for analyzing and understanding complex events, which includes various crowd and complex events (some behaviors such as getting on and off the underground, collisions, battles, earthquake escapes, etc.). The dataset includes the largest number of poses and complex event labels, which can be used for some tasks such as human detection, pose recognition, and target tracking, etc.

Due to the diversity of the Hi-Eve dataset, the various types of models are realised differently in different tasks. In the tasks of Walking and Shooting, HRNetV2 respectively achieved the best result of 81.6% and 65.7%. In the task of Fighting, HRNet achieved the best result of 71.2%. In complex environments such as Robbery, Arson and Jump, DBayC achieves the best results compared to some SOTA models. In the end, DBayC gets a 3.8% improvement in accuracy compared to HRNetV2 and achieves the best results as show in Table 1. Combined with the Figure 4, it is concluded that DBayC is superior in complex environment.

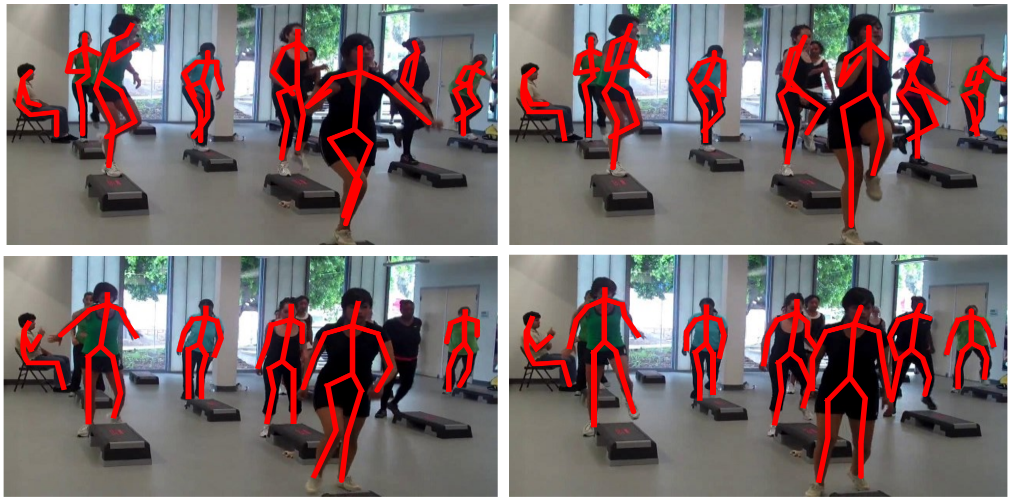

Figure 4. Pose detection effect of DBayC on Hi-Eve dataset.

Table 1. Test results of Hi-Eve dataset.

Algorithm	Year	Walking	Robbery	Fighting	Shooting	Arson	Jump	Avg
PoseNet	2018	58.6	64.9	56.7	51.9	68.2	71.3	63.7
DensePose	2018	63.7	71.8	63.5	59.7	77.9	73.6	68.2
MoveNet	2021	58.7	68.9	66.1	58.7	73.3	74.6	64.5
HRNet	2021	71.2	71.8	**71.6**	57.3	75.9	73.1	67.5
HRNetV2	2021	**81.6**	73.1	69.1	**65.7**	79.7	80.2	75.4
DBayC	**-**	76.1	**78.2**	63.7	61.2	**83.2**	**81.1**	**79.2**

3.3 Experimental comparison of PoseTrack2017 dataset

In order to demonstrate the generalizability of the DBayC algorithm, the testing was done here on the PoseTrack2017dataset and the results are shown in Table 2. PoseTrack2017 is a large dataset for human pose estimation and joint tracking in videos. It consists of 1356 video sequences, 46000 annotated video frames, and 276000 annotations of human poses.

Table 2. Results in the PoseTrack2017 dataset.

Algorithm	**Year**	**Shou**	**Elbo**	**Wri**	**Knee**	**Ankl**	**Hip**	**Avg**
MoveNet	2021	73.7	69.8	62.7	69.8	71.2	**75.8**	66.9
HRNet	2021	58.7	62.6	65.1	68.7	70.6	71.6	67.3
HRNetV2	2021	73.3	62.8	**78.6**	63.3	70.9	72.1	71.7
DBayC	**-**	**77.6**	**72.7**	71.8	**73.2**	**79.1**	75.2	78.6

As shown in Table 2, MoveNet achieved the best result of 75.8% in task of Hip. In the detection task of Wri, HRNetV2 achieved the best result of 78.6%. In the detection tasks of Shou, Elbo, Knee, Ankle, DBayC respectively achieved the best value of 77.6%, 72.7%, 73.2%, and 79.1%. Ultimately, DBayC obtains a 6.9% improvement in accuracy compared to HRNetV2 and achieves the best results.

Combining the experimental results in Tables 1 and 2, and the detection effect in Figure 4, it can be concluded that the DBayC algorithm is superior to the SOTAs algorithm. By comparing two different data sets, it is demonstrated that the DBayC algorithm has a certain degree of generalization.

4. CONCLUSION

This paper aims to solve the common problems of pose estimation algorithms based on CNN, such as low accuracy due to complex background or small detection targets. The paper designs the DBayC algorithm. First, the LBN module is designed based on the CNN algorithm to achieve the semantic segmentation effect of the human body. Then, a corresponding 15-node graph is constructed for the segmented human body. Finally, Bayesian theory is used to analyze the node relationships in the graph. Through training on two data sets Hi-Eve and PoseTrack2017, DBayC is compared with some SOTAs models. From the experimental results, we can intuitively see that the DBayC algorithm is more superior.

ACKNOWLEDGMENT

This paper is supported by Guangdong Provincial Education Department Youth Innovation Project under the Grant No.2022KQNCX118 and Guangdong University of Science and Technology No.GKY-2022KYYBK-16.

REFERENCES

[1] Campbell, D., Liu, L. and Stephen, G., "Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization," Computer Vision-ECCV 2020: 16th European Conference, (2020).

[2] Cootes, T. F., Taylor, C. J., Cooper, D. H., et al., "Active shape models-their training and application," Computer Vision and Image Understanding 61(1), 38-59 (1995).

[3] Bay, H., Tuytelaars, T. and Van Gool, L., "Surf: Speeded up robust features," Computer Vision-ECCV 2006: 9th European Conference on Computer Vision, 404-417 (2006).

[4] Toshev, A. and Szegedy, C., "Deeppose: Human pose estimation via deep neural networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1653-1660 (2014).

[5] Zheng, C., Wu, W., Chen, C., et al., "Deep learning-based human pose estimation: A survey," ACM Computing Surveys 56(1), 1-37 (2023).

[6] Qiao, S., Wang, Y. and Li, J., "Real-time human gesture grading based on OpenPose," 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 1-6 (2017).

[7] Cao, Z., Simon, T., Wei, S. E., et al., "Realtime multi-person 2d pose estimation using part affinity fields," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7291-7299 (2017).

[8] Nakai, M., Tsunoda, Y., Hayashi, H., et al., "Prediction of basketball free throw shooting by openpose," New Frontiers in Artificial Intelligence: JSAI-isAI 2018 Workshops, JURISIN, AI-Biz, SKL, LENLS, IDAA, 435-446 (2019).

[9] Wang, C. Y., Bochkovskiy, A. and Liao, H. Y. M., "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7464-7475 (2023).

[10] Hu, J., Shi, C. J. R. and Zhang, J., "Saliency-based YOLO for single target detection," Knowledge and Information Systems 63(3), 717-732 (2021).

[11] Zhang, J., Huang, M., et al., "A real-time chinese traffic sign detection algorithm based on modified YOLOv2," Algorithms 10(4), 127-139 (2017).

[12] Mao, Q. C., Sun, H. M., Liu, Y. B., et al., "Mini-YOLOv3: real-time object detector for embedded applications," IEEE Access 7, 133529-133538 (2019).

[13] Hu, X., Liu, Y., Zhao, Z., et al., "Real-time detection of uneaten feed pellets in underwater images for aquaculture using an improved YOLO-V4 network," Computers and Electronics in Agriculture 185, 106135 (2021).

[14] Kim, J. H., Kim, N., Park, Y. W., et al., "Object detection and classification based on YOLO-V5 with improved maritime dataset," Journal of Marine Science and Engineering 10(3), 377 (2022).

[15] Norkobil Saydirasulovich, S., Abdusalomov, A., Jamil, M. K., et al., "A YOLOv6-based improved fire detection approach for smart city environments," Sensors 23(6), 3161 (2023).

[16] Hussain, M., "YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection," Machines 11(7), 677 (2023).

[17] Yang, Z., Yu, X. and Yang, Y., "Dsc-posenet: Learning 6dof object pose estimation via dual-scale consistency," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3907-3916 (2021).

[18] Bajpai, R. and Joshi, D., "Movenet: A deep neural network for joint profile prediction across variable walking speeds and slopes," IEEE Transactions on Instrumentation and Measurement 70, 1-11 (2021).

[19] Sun, K., 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, 5693-5703 (2019).

[20] Wang, J., Sun, K., Cheng, T., et al., "Deep high-resolution representation learning for visual recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence 43(10), 3349-3364 (2020).

[21] Andriluka, M., Iqbal, U., Insafutdinov, E., et al., "Posetrack: A benchmark for human pose estimation and tracking," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5167-517 (2018).