Modern Management based on Big Data III
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA220118

Lightweight Gesture Pose Estimation Based on CPM Algorithm

He WANG[a,1], Qingjiang YANG[a], Qiang WANG[b] and Liang YU[a]
[a] *Heilongjiang University of Science and Technology, College of Electronic and Information Engineering, Harbin 150022, China*
[b] *Yanshan University, School of Vehicle and Energy, Qinhuangdao 066000, China*

Abstract. With the experiential enhancement of artificial intelligence products, gesture estimation, as a classic computer vision task, has a wide range of application scenarios. Aiming at the current network model that needs to be lightweight in mobile smart products, this paper designs a lightweight gesture pose estimation model based on the CPM (Convolutional Pose Machine) multi-stage human pose estimation network. A comparative experiment based on the RHD open-source data set was conducted to compare and analyze the lightweight CPM gesture estimation model while ensuring accuracy while effectively reducing the amount of model parameters, which provides a basis for the development of real-time mobile terminal gesture pose estimation.

Keywords. Convolutional pose machine, gesture pose estimation, lightweight

1. Introduction

Pose estimation is to connect joint points to judge the state and behavior of human body parts [1]. Two methods are included by Pose estimation, which are top-down [2] and bottom-up [3]. And meanwhile, the gesture posture, as an important part of human posture, is a way to accurately locate the positional relationship between the key points of the hand, and then to infer the corresponding posture estimation method. According to research methods, gesture pose estimation can be divided into three methods: Generative Methods [4], Discriminative Methods [5] and Discriminative methods and Generative methods [6]. The generation method is to match the input image with a pre-defined hand model. The discriminant method is a method of directly positioning the hand joints based on the appearance of the input depth image. Random forest [7] is a more traditional discrimination method, which can quickly and accurately process a large number of input variables, but for the shortcomings of self-occlusion and low resolution of the hand depth map, there are a lot of errors in its results. The processing method of gesture estimation using convolutional neural network is constantly overcoming these problems [8-9]. Tompson et al. [10] used CNN for the first time to locate the key points of the hand and estimate the 2D hot soil of each joint point for estimation. Ge et al. [11] directly converted the 2D input depth map into estimated 3D coordinates through 3D CNN. The hybrid method, which combines the generation method and the discriminant method, uses the discriminant method first, and then corrects the result by the generation

[1] Corresponding Author, He WANG, Heilongjiang University of Science and Technology, No. 2468, Puyuan Road, Songbei District, Harbin, Heilongjiang; E-mail: 2250293019@qq.com.

method. Ye et al. [12] used a hybrid method to estimate gestures in a multi-stage and hierarchical manner, which has good robustness. Hwang, J et al. [13] used a simple K-means clustering algorithm to explore unusual poses that are rare which occupy a small portion in a pose dataset. B Artacho et al. [14] proposed OmniPose, a single-pass, end-to-end trainable framework, that achieves state-of-the-art results for multi-person pose estimation. Hietanen, A et al. [15] introduce an approach that connects error in pose and success in robot manipulation, and propose a probabilistic performance measure of the task success rate.

We are in the era of the popularization of big data, and electronic products such as artificial intelligence have emerged one after another, and have achieved remarkable development. With the introduce of deep learning, more advanced and novel methods such as human body gesture recognition and gesture estimation in the field of human-computer interaction have relatively broad application prospects. Of them, gesture estimation based on machine vision, as one of the core technologies of human-computer interaction, has made good progress. In summary, posture estimation of gestures is a task that involves the intersection of natural language processing, pattern recognition, and computer vision. The effective estimation and recognition of gestures has promoted the advancement of machine learning and other fields, and has accelerated the process of globalization, so it has certain research significance.

With the development of computer vision, gesture estimation is often used in many scenarios such as human-computer interaction and video surveillance. Generally, convolutional neural networks have a large amount of parameters and calculations, which cannot be applied to mobile terminals or embedded devices. In order to better implement real-time gesture pose estimation on mobile terminals or embedded devices, the network model needs to be lightweight. Based on the multi-stage convolutional pose machine CPM (Convolutional Pose Machine) network in human pose estimation [16], this paper designs the gesture pose estimation network. At the same time, a lightweight network is designed to ensure the accuracy and stability. The above reduces the parameter amount of the model.

2. CPM Algorithm

The Convolutional Pose Machines (CPM) algorithm is a multi-stage pose estimation network based on serialized fully convolutional network structure and learning spatial information. It integrates the convolutional network into Pose Machines to learn image features and image-related spatial models; a serialized structure composed of a full convolutional network. The convolutional network is directly operated on the belief maps of the previous stage. Output more refined joint point position estimation results to handle structured prediction tasks; an end-to-end learning network that uses intermediate supervision loss to solve the problem of gradient disappearance. As shown in Figure 1, the CPM algorithm is a multi-layered network structure composed of a multi-stage convolutional neural network. It generates a confidence map through the convolutional neural network of each stage, and then predicts the position of each key point in each stage. Among them, gt represents the convolutional neural network of each stage, bt represents the confidence map, and uses t to represent each stage in the multi-level sequential structure, and when t>1 becomes the strengthening stage.

For the initial stage, the feature extraction network is used to extract features from the image, and the position confidence of each location is predicted by g_t. For the

enhancement phase, the CPM network uses a feature extraction network to re-extract features from the original image, and map and merge it with the confidence map of which phase. Through the phase loss function value obtained at the end of each phase, the CPM algorithm is locally supervised.

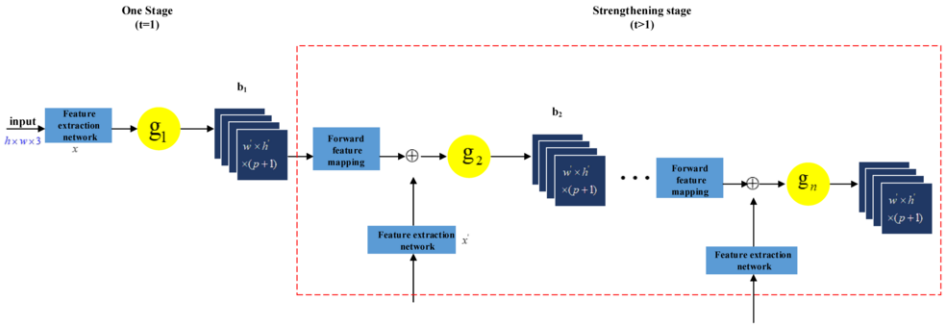

Figure 1. Convolutional attitude machine CPM network diagram.

At the same time, the CPM algorithm has the disadvantage that a large amount of computing power is required to extract the original image features each time because the feature extraction network is not unified. In addition, CPM needs to perform forward mapping features during the enhancement phase, which consumes a lot of computing power and is more complicated.

3. CPM Lightweight Gesture Pose Estimation Algorithm

CPM (Convolutional Pose Machine) pose estimation network is often used for human pose detection. Based on the idea of CPM multi-level sequential pose estimation network, this paper will design the network structure and apply it to gesture pose estimation. In view of the huge amount of calculation in the process of extracting features by the CPM algorithm, this paper changes part of the standard convolutional layer structure to a deep separable convolutional layer, and merges the losses in each stage to obtain a lightweight CPM gesture pose estimation network. It provides a basis for real-time estimation and detection of gesture estimation on mobile or embedded devices.

3.1. CPM Gesture Network Pose Estimation Structure

The structure composed of the convolution structure and the attitude machine structure is called the convolution attitude machine [17]. This structure can automatically learn features from the training data set, and inferentially learn the distance structure relationship between key points, which can be applied the location of the key points in the gesture pose estimation in this paper.

The process of the migrated CPM hand posture estimation network to return the coordinate position information of the key nodes of the image is: use the response heat map heatmap and feature map representing the spatial constraint information between the various joint points as data to be transmitted in the network; use the multi-stage convolutional neural network conducts supervised training and processes the response information of key nodes. The detailed structure of the algorithm is shown in Figure 2.

From the structure diagram of the CPM gesture posture estimation network, it can be seen that the network is performing the overall process of hand posture estimation. Among them, the convolution part of the CPM pose estimation network includes

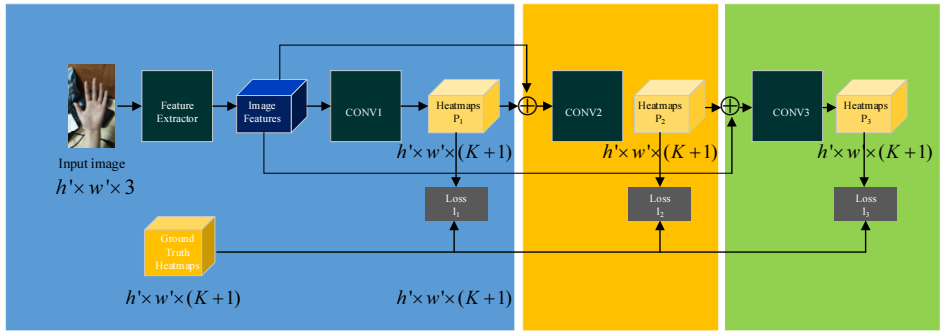

Figure 2. CPM gesture posture network structure diagram.

convolution and maximum pooling layers. The network modified based on the VGGNet network is used for backbone feature extraction network features. In the figure, CONVx (x=1,2,3) is composed of a series of convolutional layers and pooling layers.

The CPM pose estimation network is a multi-stage prediction network. Based on the topological relationship between the key points of gestures, the accuracy requirements of gesture pose estimation can be met by using a lower stage. Therefore, the CPM gesture pose estimation network used in this paper contains a total of three stages. The heat map of each stage is used to characterize the position information of the key nodes, and the stage corresponding to the smaller receptive field is located in the previous stage of the network, so the relative error of the obtained prediction result is relatively large. The later stage of the enhancement layer corresponds to a larger receptive field, and the context information and image features of the previous stage can be obtained, so the prediction result with relatively small error is obtained. This staged posture estimation network can make a step-by-step more accurate reasoning of the hand posture, and finally fuse the obtained loss values to obtain the final predicted joint point position information.

In view of the mobile terminal or embedded pose estimation task, a lighter model is required. In order to reduce the overall parameter amount of the model, the convolution kernel in the three-stage CPM gesture pose estimation network is replaced with 1*1 and 3*3 convolution kernel. This method can more accurately estimate the key point positions and reduce the amount of model parameters, and obtain a lighter gesture pose estimation network.

3.2. The Introduction of Deep Separable Convolutional Layers

Depth separable convolution is a kind of factorized convolution realized by decomposing ordinary convolution into two parts of deep convolution and 1*1 point-by-point convolution. In the first step, the deep convolution operation uses a single convolution and a lightweight filtering operation on the input channel, that is, to realize the convolution operation of the convolution kernel and the feature map one by one. In the second step, a 1*1 convolution kernel is used to perform a point-by-point convolution operation on the basis of the previous feature map, and the output feature maps are linearly combined. Compared with the one-step operation of the ordinary convolutional layer, the depth separable convolution uses two steps of deep convolution and point-by-point convolution to convolve a single channel and then combine them. The parameter amount and calculation amount of the model can be greatly reduced. Among them, the comparison structure diagram of the standard convolutional layer and the depth separable convolutional layer is shown in Figure 3.

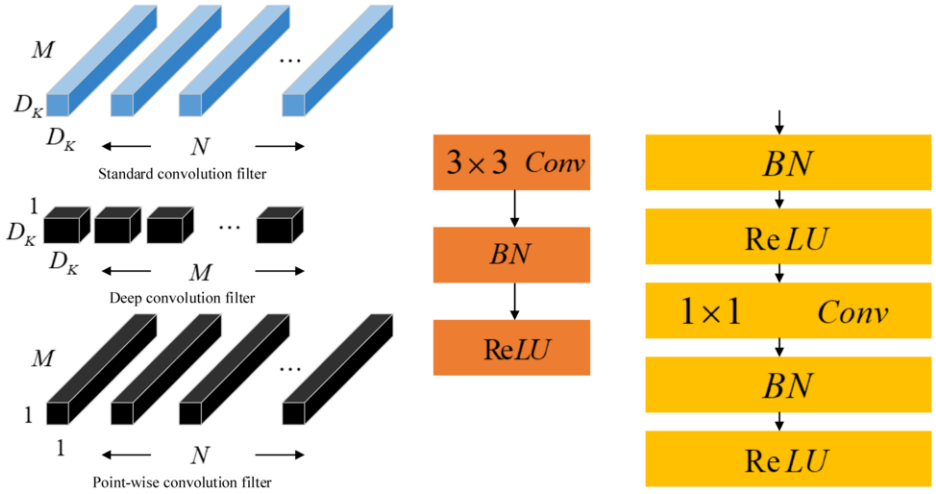

Figure 3. Comparison of ordinary convolutional layer and depth separable convolutional layer.

Among them, the calculation amount and parameter amount of the depth separable convolution and the standard convolution layer are compared as follows [18]:

Calculation amount of standard convolution:

$$D_K \times D_K \times N \times M \times D_F \times D_F \tag{1}$$

Parameters of standard convolution:

$$D_K \times D_K \times N \times M \tag{2}$$

The amount of calculation for deep convolution:

$$D_K \times D_K \times 1 \times M \times D_F \times D_F \tag{3}$$

The amount of calculation for point-by-point convolution:

$$1 \times 1 \times N \times M \times D_F \times D_F \tag{4}$$

The parameters of the depth convolution:

$$\mathrm{D}_K \times \mathrm{D}_K \times 1 \times M \tag{5}$$

The number of parameters for point-by-point convolution:

$$1 \times 1 \times N \times M \tag{6}$$

The parameters of the depth separable convolution:

$$D_K \times D_K \times M + N \times M \tag{7}$$

The amount of calculation for the depth separable convolution:

$$D_K \times D_K \times M \times D_F \times D_F + N \times M \times D_F \times D_F \tag{8}$$

The above formula shows that deep separable convolution can greatly reduce the amount of network parameters and calculations. Therefore, this paper introduces deep separable convolution to replace part of the convolutional layer structure in the CPM multi-stage attitude estimation network, resulting in a lighter weight CPM gesture pose estimation network.

4. Experiments and Results

4.1. Experimental Environment and Data Set

The environment used in this paper is anaconda+keras2.24+TensorFlow1.14+cuda9.2, graphics card NVDIAGTX2080.

This paper uses the RHD [19] open source gesture and pose dataset as the training dataset. This dataset is a gesture dataset of synthetic RGB images composed of 41258 training images and 2728 test images. It is obtained by asking 20 different mannequins to perform 39 different actions randomly and then generating arbitrary ones after all. Due to the huge changes in the viewpoint and hand ratio, and the huge visual differences caused by the random noise and ambiguity of the image, the data set is quite challenging. For each RGB image, it provides the corresponding depth image, occlusion label, 2D label and 21 key point 3D label. Figure 4 shows part of the dataset

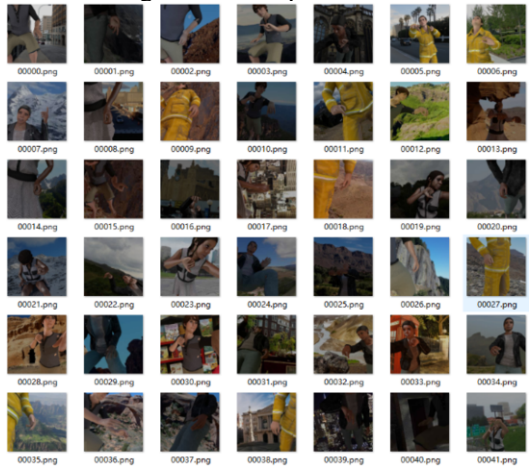

Figure 4. Part of the RHD dataset.

4.2. Training and Evaluation Indicators

We use the RHD data set to divide it into training set, test set, and validation set with a ratio of 8:1:1 for training. Set the Batch size to 8, the initial learning rate is 0.0001, the

number of iterations is 60,000, and the learning rate drops to 10 times at each iteration of 10,000, and the Adam optimization method is used.

4.2.1. Test Set Evaluation Model

This paper uses model accuracy and model size to measure the relative performance of the gesture pose estimation model proposed in this paper.

Model accuracy: PCK (Percentage of Correct Keypoints) means that the normalized distance between the 21 actual keypoints of the gesture and the predicted keypoint is less than the percentage of the specified threshold. Taking the threshold as the abscissa and the PCK as the ordinate, the area under the PCK curve under different error thresholds is drawn as AUC (Area Under Curve, AUC) [20].

As shown in Table 1, compared with the original model performance indicators without lightweight processing, experiments show that when the model accuracy is stable, the size of the model is greatly reduced (here the Loss function is).

Table 1. Lightweight model comparison experiment.

Model	Model accuracy (%)	Model size (M)	Flops
CPM	72.8	24.32	48.62
Lightweight CPM	72.4	6.89	13.77

It can be seen from the table that compared to the original CPM gesture posture estimation network, the introduction of a deeply separable network model for lightweight processing of CPM gesture posture can reduce the model size to about 4 times the original under the premise of ensuring the stability of the model. The Flops value will be about 1/4 times the original value. It can effectively reduce the amount of model parameters and provide a basis for the development of real-time mobile terminal gesture pose estimation.

4.2.2. Visualization of Detection Results

Figure 5 shows the effect of the detection effect on the RHD dataset under the anaconda+keras2.24+TensorFlow1.14+cuda9.2 experimental environment.

Figure 5. Detection effect visualization.

5. Conclusion

With the development of computer vision, gesture estimation is often used in many scenarios such as human-computer interaction and video surveillance. Generally, convolutional neural networks have a large amount of parameters and calculations, which cannot be applied to mobile terminals or embedded devices. In order to better implement real-time gesture pose estimation on mobile terminals or embedded devices, the network model needs to be lightweight. This paper changes the network structure of CPM based on human pose estimation and transfers it to gesture pose estimation. At the same time, the introduction of deep separable convolution to replace part of the convolutional layer structure in the network structure greatly reduces the amount of calculation and calculation of the model.

Experimental comparison and visualization show that the lightweight CPM attitude estimation network can still have good detection results in complex environments such as occlusion and deformation, and can greatly reduce the amount of model parameters and parameters while ensuring stable accuracy. The amount of calculation lays the foundation for real-time running of gesture and posture algorithms on mobile or embedded devices.

References

[1] Ramakrishna V, Munoz D, Hebert M, Bagnell JA, Sheikh Y. Pose machines: articulated pose estimation via inference machines. Lecture Notes in Computer Science. 2014; 33-47.
[2] Cao Z, Hidalgo G, Simon T, Wei S E, Sheikh Y. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018; 1.
[3] Fang HS, Xie S, Tai YW, Lu C. RMPE: regional multi-person pose estimation, 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017:2353-2362, doi: 10.1109/ICCV.2017.256.
[4] Qian C, Xiao S, Wei Y, Tang X, Jian S. Realtime and robust hand tracking from depth. 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014 :1106-1113.
[5] Ge L, Hui L, Yuan J, Thalmann D. Robust 3D hand pose estimation in single depth images: from Single-View CNN to Multi-View CNNs. 2016: 3593-3601.
[6] Ye Q, Yuan S, Kim T K. Spatial attention deep net with partial PSO for hierarchical hybrid hand pose estimation. 2016:346-361.
[7] Keskin C, Kıraç F, Kara YE, Akarun L. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. 2012:852-863.
[8] Liang H, Yuan J, Thalmann D. Parsing the hand in depth images. IEEE Transactions on Multimedia, 2014, 16(5):1241-1253.
[9] Chen X, Wang G, Guo H, Zhang C. Pose guided structured region ensemble network for cascaded hand pose estimation. Neurocomputing, 2020: 395, 138-149.
[10] Tompson J, Stein M, Lecun Y, Perlin K. Real-Time continuous pose recovery of human hands using convolutional networks. ACM, 2014:1-10.
[11] Ge L, Liang H, Yuan J, Thalmann D. 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 1991-2000.
[12] Ye Q, Yuan S, Kim T K. Spatial attention deep net with partial PSO for hierarchical hybrid hand pose estimation. 2016: 346-361.
[13] Hwang J, Yang J, Wak NK. Exploring rare pose in human pose estimation. IEEE Access, 2020, 8:194964-194977.
[14] Artacho B, Savakis A. OmniPose: A multi-scale framework for multi-person pose estimation. 2021. arXiv preprint arXiv:2103.10180
[15] Hietanen A, Latokartano J, Foi A, Pieters R, Kmrinen JK. Benchmarking pose estimation for robot manipulation. Robotics and Autonomous Systems, 2021, 143(1):103810.

[16] Liu Z, Cai YF, Wang H, Chen L, Gao HB, Jia YY, Li YC, Robust Target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather conditions, in IEEE Transactions on Intelligent Transportation Systems, doi: 10.1109/TITS.2021.3059674.
[17] Ramakrishna V, Munoz D, Hebert M, Bagnell JA, Sheikh Y. Pose machines: articulated pose estimation via inference machines. European Conference on Computer Vision. Springer, Cham, 2014: 33-47.
[18] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H, MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.
[19] Zimmermann C, Brox T. Learning to estimate 3D hand pose from single RGB images. 2017: 4903-4911.
[20] Zhao, B, Zhao, B, Tang, L, Wang, W, Wu, C. Multi-scale object detection by top-down and bottom-up feature pyramid network. System Engineering and Electronic Technology (English Version), 2019, 30(1): 1-12.