Multi person pose estimation based on improved openpose model

Jiayuan Xing1*, Jun Zhang1 and Chenxing Xue1
1School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin, China

*Corresponding author e-mail: 18822148155@163.com

Abstract. In order to solve the problem of multi person pose estimation, a human body pose estimation model framework based on openpose from American Carnegie Mellon University is presented, which uses the COCO (Microsoft common objects in context) data set to augment the data, optimize the network structure and retrain, adjust the learning rate properly and the convergence speed in the early stage of the model, finally using the model to carry out research on human key point detection and visualizing analysis in a multi-person complex scene. It is found that the improved openpose model adopts a bottom to up detection strategy to avoid the influence of the number of human bodies on real-time performance, and has higher key point detection efficiency than the top to down method. The study provides a real-time position feedback scheme for human pose estimation in complex scenes.

1. Introduction

Human pose estimation has always been the key direction of computer vision development. The human body pose estimation is to identify the position of the key points of the human body from a single RGB image or video and construct a human skeleton model. From the perspective of historical development, there are two major schemes in general, one is to use the pictorial structures method, and the other is based on the deep learning method. The component matching method [1, 2, 3] mostly uses SIFT [4], HOG [5] to extract features, and then uses SVM [6] classifier to complete classification. Most of the deep learning is designed as an end-to-end network, and the mapping relationship between the annotation information and the real key points is learned through the powerful fitting ability of the neural network. In addition, for the multi-person pose estimation problem, the deep learning method is divided into two framework ideas: top to down [7, 8, 9] and bottom to up [10]. The difference is that the former first crops the single person of the input image, and then uses the single pose estimation method to complete the estimation. The latter first detects all the key points in the input image and then clusters into individual human bodies by the combination method. The former has the advantage of high precision, and the latter has the characteristics of real-time.

In this study, the multi-person pose estimation algorithm is further improved by increasing the robustness of the model and improving the convergence speed by using the bottom-up approach and the openpose model as the framework. In the training process, the COCO [11] data set is used to extract the training samples and annotation information of the human key points and perform visual analysis, and finally fed to the openpose model for training. In the image testing phase, the relationship between the key points heat map and the limb connection was analyzed using the trained model.
2. Openpose algorithm improvement

The CPM [12] proposed in 2016 is a representative of the bottom to up method, and the network model is very robust. CPM uses a sequential convolutional architecture to express spatial and texture information. The openpose convolutional neural network is based on the improvement of CPM. The network architecture is shown in Figure 1.

The improved openpose network model has two branches for the detection of key points and the combination of key points. At the same time, the model cleverly designs the front features to participate in the subsequent convolution operation, retains the key point feature information, and designs the relay supervision intermediate supervision method for the problem of gradient vanish to ensure that the back layer gradient can be transferred to the shallow network to change the convolution core weight value, so as to achieve the optimization purpose of reducing the distance between the predicted heat map of the key point of the model and the coordinate value of the real annotation information. Using resnet18[13] network instead of vgg-19 network to extract image features, the blue block represents the upper branch, which is used to extract the position of human key points; the green block represents the lower branch, which is used to integrate the position information of key points and realize the correct connection between key points. The residual idea of the resnet18 network solves the gradient problem and improves the performance of the overall network.

2.1. Key point detection branch

The upper branch takes the acquired features as input and outputs a human bone key point confidence network. In order to solve the problem of inaccurate key point and difficulty in distinguishing left and right key points, the multi-stage strategy was used to train the convolution network. We designed the network into three stages for the purpose of reducing the amount of computation. The convolution kernel of stage one is kept 3x3 for better detection of large objects. In the subsequent stage, the 7x7 convolution kernel is modified to 5x5, ignoring the small-scale key points, and sacrificing a certain precision in exchange for the real-time improvement of the model.

2.2. Key point assemble branch

The lower branch also uses image features as input, performs multiple convolutions, and outputs the partial affinity field(PAF) of human bone key points. After three stages, the key point confidence maps set and the key points partial affinity filed set work together to realize the multi person pose estimation function. The final output of the upper branch is the set $s = (S_1, S_2, ... S_n)$, $S_n$ is the...
The confidence of the nth key point of the human body. The final output of the lower branch is the set 
\( L = (L_1, L_2, \ldots, L_n) \), \( L_n \) is the nth connection affinity of key points in human body.

\[
S_{j,k}^*(p) = \exp\left(\frac{\|P - x_{j,k}\|^2}{\sigma^2}\right) 
\]  

(1)

\[
S_j^*(p) = \max_k S_{j,k}^*(p) 
\]  

(2)

\[
f_j^* = \sum_{j=1}S_j^*(p) - S_j^*(p) 
\]  

(3)

The set \( S \) represents the probability that the predicted key point location of the network is the real 
position. Formula for single-person k confidence \( S_j (1) \), \( x_{j,k} \) represents the labeled position of the jth 
key point of the kth person in the training image, P is the coordinate point of the jth key point 
predicted by the network, \( \sigma \) stands for standard deviation and the "*" represents stage. Formula for 
multi-person confidence \( S_j (2) \), take the maximum of multiple standard deviations as the peak of the 
key point, take the maximum of the multiple normal distributions as the predicted position of the key 
point. In formula (3), the RMSE (Root Mean Squared Error) between the real coordinate value and the 
model prediction value is used to define the loss value of the upper branch. Where t is the ground truth 
of j key point.

3. Experimental process

3.1. Data preparation

At present, Ms coco data set is a rich data set to solve the task of target detection and recognition, 
semantic segmentation, in which the key point annotation information is used for pose estimation. 
After the data set is prepared, the annotation information stored in JSON format can be accessed 
efficiently by using cocoapi. The specific steps are as follows: First, convert the original annotation 
file ‘.json’ format to ‘.mat’ format for the convenience of subsequent processing on Matlab. The 
original json annotation content is shown in Figure 2. For example, an image id is 855. There are 8 
people in the picture, and 3 of them are marked with key points. Use cocoapi to display key points and 
segmentation information, as shown in Figure 3. Second, generate mask template to distinguish human 
key points (foreground) from other things (background). Third, the data format is converted back to 
the ‘.json’ format. Finally generate LMDB data form.

Before the training, the training set is expanded by data aumentation [14] techniques such as image 
horizontal flipping, zooming in or out, horizontal offset, vertical offset and rotation by angle.

![Person_keypoints_val2017.json stores information](image.png)
3.2. Environmental preparation

System construction requires two aspects of hardware and software. Hardware: CPU intel® CoreTM i5-4590, GPU GeForce GTX 980 4G; Software: Ubuntu 16.04.4 LTS, CUDA 9.0.176, Cudnn 7.3.0, MATLAB (R2016b), caffe and caffe_train_master versions.

3.3. Model training

Load the resnet18 model parameters as the feature extractor, modify the openpose model training learning rate is 0.0008, use the momentum back propagation optimization function, the initial value is set to 0.9, the maximum number of iterations is 10,000 steps, and the training batch size batch_size is set to 16.

Taking the first stage as an example, the change in the loss value of the training process is shown in Figure 4. Where L1 represents the loss value of the key point detection, and L2 represents the loss value of the correct combination between the key points.

3.4. Experimental results and analysis

There are 18 key points marked by coco data set, including one background. In addition, openpost adds an additional neck key point to increase the contextual relevance between the head and upper body. Taking the multi person still image as the test target, the results are shown in Figure 5.
Compared with the original openpose original 0.0004 learning rate, it is found that an appropriate increase in the learning rate to 0.0008 can quickly reduce the loss value of the previous period. The influence of different learning rates on the model is shown in Table 1.

Table 1. The influence of different learning rates on the model.

| Different models | Still image | 10000 iteration-L1 loss | 10000 iteration-L2 loss | 10000 iteration-total loss |
|------------------|-------------|-------------------------|-------------------------|----------------------------|
| Learning rate 0.0004 | 152.682 | 46.6257 | 1086.52 |
| Learning rate 0.0008 | 118.792 | 39.5757 | 856.76 |

Only using the three-stage openpose model, the parameter quantity is reduced, and the FPS has a certain improvement compared with the original 6-stage model, as shown in Figure 6. At the same time, it can be found that the detection efficiency does not decrease with the increase of the number of people, which is the advantage of down to top.
4. Conclusion

Human pose estimation is an important task in the field of computer vision. Planar multi-person pose estimation is the premise of complex 3D pose estimation. Deep learning extracts the abstract features of key points through a multi-layer convolutional layer, fitting the relationship between the label position and the training image. This study optimized the openpose depth learning pose estimation method, analyzed the annotation form of COCO dataset, used data augmentation and retraining, and used datasets with relevant key points challenge questions to study different learning rates in the model training process. The impact of verifying the appropriate increase in learning rates helps speed up the model's testing cycle.

Improved openpose model adopts the bottom-up approach to solve the problem that the traditional scheme has large calculation and poor real-time performance in multi-person scenarios, but it also fails to make full use of the strong context of the human body itself. In the future work, the human body features will be further explored and integrated into the deep learning method.

References

[1] MA Fischler and RA Elschlager. The representation and matching of pictorial structures. Computers, IEEE Transactions on, 100(1):67–92, 1973.

[2] O’ROURKE J, BADLER N I. Model-based image analysis of human motion using constraint propagation[J]. IEEE Trans Pami, 1980, 2(6): 522-536.

[3] Felzenszwalb and D. Huttenlocher, “Pictorial structures for object recognition,” International Journal of Computer Vision, vol. 61, no. 1, pp. 55–79, 2005.

[4] Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International journal of computer vision, 2004, 60(2):91-110.

[5] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. computer vision and pattern recognition, 2005: 886-893.

[6] Cortes C, Vapnik V. Support Vector Network[J]. Machine Learning, 1995, 20(3):273-297.

[7] Insafutdinov E, Pishchulin L, Andres B, et al. DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model[C]. european conference on computer vision, 2016: 34-50.

[8] He K, Gkioxari G, Dollar P, et al. Mask R-CNN[C]. international conference on computer vision, 2017: 2980-2988.

[9] Newell A, Yang K, Deng J, et al. Stacked Hourglass Networks for Human Pose Estimation[J]. arXiv: Computer Vision and Pattern Recognition, 2016.

[10] Cao Z, Simon T, Wei S, et al. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields[C]. computer vision and pattern recognition, 2017: 1302-1310.

[11] Lin T, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context[C]. european conference on computer vision, 2014: 740-755.

[12] Wei S, Ramakrishna V, Kanade T, et al. Convolutional Pose Machines[C]. computer vision and pattern recognition, 2016: 4724-4732.

[13] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]. computer vision and pattern recognition, 2016: 770-778.

[14] Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation[J]. 2017.