Milestone for Pose Estimation and Application in Fitness

Song Jiafei, Wang Xin
ShanghaiTech University
{songjf, wangxin1}@shanghaitech.edu.cn

Abstract

In this project, we first survey on various implementations of human pose estimation both for single and multiple people.We will focus more on single-person one. Learning the structure of the network and the quintessence of each paper is quiet important in current stage. Based some implementation, we will then do some modification and improvement for better application in fitness. How different your actions are from your fitness coach's actions is difficult to determined by yourself.However, the development of pose estimation can point out such differences.

1. Introduction

Human 2D pose estimation is a popular and fundamental topic in computer vision. Given a single RGB image, we wish to determine the precise pixel location of important keypoints of the body. Knowing the body structure of human can be beneficial for further high-level computer vision tasks, such as abnormal behavior detection and so on. What's more, there are a lot of applications of pose estimation in human-computer interaction and one famous example is Kinect which is produced by Microsoft for Xbox 360 and Xbox One video game consoles.

And in this project, we will apply pose estimation to the fitness. In strength training and fitness, for example, the squat is a compound, full body exercise that trains primarily the muscles of the thighs, hips and buttocks. You should be very careful about such exercises, wrong actions will do lots of harm to your body but improper actions have no effect on your muscles. Hence, if we can record the pose of your coaches and yours, it's more convenient and easy to adjust your actions in daily training.

1.1. Datasets

MPII [1] and COCO [2] are two famous datasets for both single-person and multi-person pose estimations. The MPII dataset includes around 25K images containing over 40K people with annotated body joints. The COCO dataset contains more than 200,000 images and 250,000 person instances labeled with keypoints (the majority of people in COCO at medium and large scales). In this project, we will train all the network with such datasets. The training time is so long. For example, after resizeing the image in COCO dataset to 256x192 from 480x640 , it will take up to 2 days to train with a pre-trained ResNet with some modification.

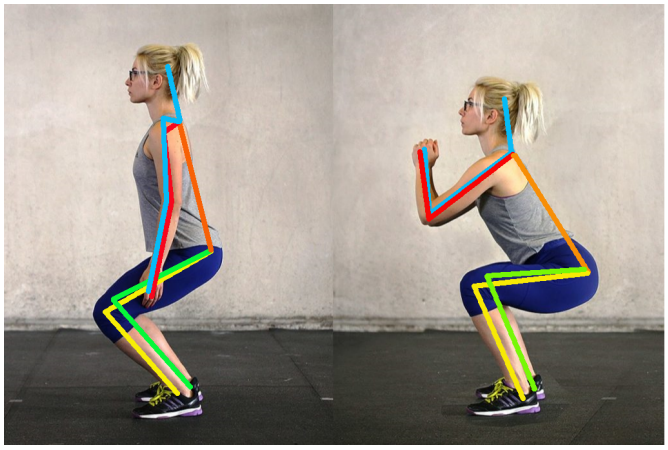

Figure 1. Pose estimation of squatting.

1.1.1 Evaluation of MPII

PCKh: PCK [3] measure that uses the matching threshold as 50% of the head segment length.

1.1.2 Evaluation of COCO

Using AP(Average Precision) and AR(Average call) to measure the performance of algorithms. Also defining the object keypoint similarity (OKS) which plays the same role as the IoU (Intersection over Union). The specific definition of OKS won't be shown here. Besides these methods, this benchmark also uses up to 10 metrics to measure the performance. These metrics are different AP/AR in different OKS.

1.2. Single-person Pose Estimation

In this part, we will show some paper we read and some related works. Since the method of deep learning has gained so much good performance, some traditional [4] methods of pose estimation have been gradually replaced.Recent pose estimation systems have universally adopted ConvNets as their main building block to replace the hand-crafted features and graphical models.

Newell [5] proposed a very interesting network, named Stacked Hourglass Networks. It allows for repeated bottom-up, top-down inference across scales. Also, it used residual module to deal with the vanish or gradient in such deep networks.

Based on such Stacked Hourglass model, Xiao Chu [6] adds multi-context attention mechanism both globally and locally to it. Attention model is quiet reasonable here and achieves some improvements but only a little.

Megvii[7] done by Face++ is the leading team in COCO Keypoint Challenge17. They proposed a very complicated network to integrate high resolution information from the earlier layers and semantic information from the deeper layers.

One more exciting implementation of single-person pose estimation is [8] done by MSRA. They ranks second after Face++.Their network is quiet simple and only contains a pre-trained 10-layer Resnet and three deconvolution layers. It amazed me that such a simple network can achieve such remarkable performance. It inspires me that how important the features are in pose estimation.

1.3. Multi-person Pose Estimation

One bottom-up excellent work is named Openpose [9] done by CMU. The performance is quiet good. They combine the confidence maps for part and the part affiity fields to predict the heatmaps of the joint. The network is also complicated. But the idea of PAF inspires me a lot.

2. Some experiments

For milestone period, after various reading and survey, we only re-implement some network. For preference, we choose the work done by MSRA. Hence, the current work mainly focuses on it. For the limitation of GPUS, we only test the result on COCO dataset and MPII dataset once.

The model only involves the pre-tained ResNet50 network and some deconvolution layers. Considering the size of GPU memory, we use batchsize 32 and resize the original image with size 256x192. It takes 14 hours to run 80 epochs, which is half of what the paper mentions.

The optimization schedule is using adam with learning rate 1e-3.

The difference of results is quiet small with the author's implementation.

Evaluation	AP	AP .5	AP.75	AR	AR .5	AR .75
Mine	0.681	0.904	0.760	0.714	0.911	0.781
Author	0.704	0.886	0.783	0.763	0.929	0.834

Table 1. Results on COCO val2017 dataset

Head	95.481	96.351
Shoulder	94.418	95.329
Elbow	86.878	88.989
Wrist	83.112	83.176
Hip	88.230	88.420
Knee	83.86	83.960
Ankle	79.448	79.584
Mean	87.347	88.532

Table 2. Results on MPII val dataset(PCKh)

3. Technical Approach

The following schedule is divided into two parts.

The first part is improving the performance of pose estimation. Since we have implemented the backbone of the network, the next stage is to modify it. Thanks to the various survey, the experiments will focus on attention mechanism and PAFs applying in our model. Additionally, how ResNet helps pose estimation may be another point we are interested in.

The second part is applying our model to the real cases– fitness. We will record various fitness videos and process them. One difficulty of our evaluation is how to tell and determine the difference between two people's pose.

We will work hard to finish thie project.

References

[1] Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation:New benchmark and state of the art analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)

[2] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P.,Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740755. Springer (2014)

[3] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. PAMI13.

[4] Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: Computer Vision and Pattern Recognition, 2008.CVPR 2008. IEEE Conference on, IEEE (2008) 18

[5] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Eu-

ropean Conference on Computer Vision. pp. 483499. Springer(2016)

[6] Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 18311840 (2017)

[7] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J. (2017). Cascaded pyramid network for multi-person pose estimation. arXiv preprint arXiv:1711.07319.

[8] Xiao, B., Wu, H., Wei, Y. (2018). Simple Baselines for Human Pose Estimation and Tracking. arXiv preprint arXiv:1804.06208.

[9] Cao, Z., Simon, T., Wei, S. E., Sheikh, Y. (2016). Realtime multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1611.08050.