Towards Realtime 2D Pose Tracking: A Simple Online Pose Tracker

Anonymous ICCV submission

Paper ID ****

Abstract

Recently, 2D human pose estimation has become a rising topic in the computer vision field. Convolutional neural networks have been shown very effective in extracting multi-person poses in still images. However, more advanced applications such as human behavior recognition and scene analysis, would require the trajectory of human pose in a short video clip. In this paper, we propose a simple online pose tracker to associate human poses across frames after detecting them in each frame. In order to boost the efficiency of our method, we 1) train a small fully convolutional network to perform multiple person pose estimation in each frame, 2) propose a metric to directly compute the association cost of individual person from different frames, and 3) find the optimal association by solving the combinatorial optimization problem given the computed cost matrix. Our method achieves state-of-the-art performance on the PoseTrack dataset, and is capable of running at ∼25 FPS on a TITAN X Pascal GPU.

1. Introduction

Human pose estimation tries to find the position of joints (anatomical keypoints) for each person in images or videos. Pose tracking in videos not only performs pose estimation in each frame but also maintains the trajectory consistency of keypoints of each person across frames. It is of critical importance for higher-level applications such as person identification, action recognition and scene analysis. However, pose tracking in videos is challenging in twofold: firstly, pose estimation of multiple person in video frames is more difficult than still images due to motion blur, camera movement and heavy occlusions; secondly, maintaining the person identity over time to obtain person keypoints trajectories is non-trivial especially when efficiency is taken into consideration.

In this paper, we mainly focus on the efficiency of pose tracking and propose a Simple Online Pose Tracker (referred to as SOPT for short) which can perform realtime multi-person pose tracking in videos. The pipline of SOPT follows the classical tracking-after-detection paradigm. For each frame, we first extract human instance with 2D keypoints via a pre-trained convolutional neural network (CNN). So, pose tracking can be viewed as a Multiple Object Tracking (MOT) problem since each person is represented by his keypoint locations. Then, the MOT problem can be directly solved by constructing a cost matrix and finding the optimal assignment. Our proposed method does not require other features except for keypoint position. Hence, it is capable of realtime multiple person pose tracking on a TITAN X Pascal GPU.

2. Simple Online Pose Tracker

In this section, we describe the proposed method, which consists of two components, i.e. single frame pose estimation and pose association across frames.

2.1. Single Frame Pose Estimation

Since there is no public dataset concerning about pose estimation in videos, we first train a CNN on the MS-COCO dataset [3] which contains annotated human keypoints in still images. The CNN architecture basically follows CMU-Pose [1], which is a multi-stage fully convolutional network. To speedup per-frame pose estimation with moderate precision loss, we modify the model by sharing more hidden units and use less number of stages. This pre-trained model on COCO dataset is referred to as Model1.

Due to the fact that the predefined person keypoints of the PoseTrack and COCO dataset are different, we cannot directly finetune Model1 on the PoseTrack dataset. Hence, we add a new branch at the end stage of Model1 to make the model learn new keypoints: head top, head bottom and nose (the meaning of nose in PoseTrack and COCO dataset are different). Remark that some new Part Affinity Fields (PAFs) are also added to connect keypoints into human instance [1]. This finetuned model on PoseTrack dataset is referred to as Model2.

2.2. Pose Association Across Frames

As for tracking each individual person over time, we propose a metric of two person based on the 2D pose locations,

which is composed of the Euclidean Distance of two person corresponding keypoints C_{kpt}, Euclidean Distance of the limb length of two person C_{limb}, and the overlap of bounding box C_{bbox}. Therefore, the cost of matching two person i, j from different frames is given by

$$C_{i,j} = \alpha C_{kpt} + \beta C_{limb} + \gamma C_{bbox} \quad (1)$$

Then, for two consecutive frames, suppose at time $t-1$ there are M person instance and at time t there are N person instance, we can construct a cost matrix $\mathcal{C}$ given by

$$\mathcal{C} = [C_{i,j}]_{M \times N} \quad for \quad 1 \leq i \leq M, 1 \leq j \leq N \quad (2)$$

Given the cost matrix $\mathcal{C}$, we can apply the Munkres algorithm to find the optimal assignment by minimizing the total cost. Besides, we set a threshold C_{thres}, and for any matched two person, we check the matching cost and remove the match if the matching cost is greater than C_{thres}. Remark that those person instance that not matched are stored to match with future detections in order to alleviate the short occlusions and miss detections.

2.3. Miscellaneous

New target trajectory is initialized when a certain detection is associated with previous frame detection which is not assigned to any other targets. In other words, a matched pair of detection from two consecutive frames will generate a new target birth. Also, a threshold is set to guarantee there are minimal number of valid keypoints of a new born target to eliminate some false positives.

Target is determined dead when a target cannot be matched in τ consecutive frames. The parameter τ should be set properly according to the short occlusion time interval of each individual.

3. Results

3.1. Single Frame Pose Estimation

We compare the pose estimation performance of the pretrained Model1 and finetuned Model2 on the PoseTrack validation set, and the Average Precision (AP) of each joint is shown in Table 1. It is clear that the proposed finetune approach can learn new point in head and also slightly increase the AP of other joints. This is because Model2 has exactly the same capability of predicting joints shoulder/elbow/wrist/hip/knee/ankle, and also Model2 learned new head joints and new PAFs which makes the human instance more likely to be connected and detected. Besides, the training process of Model2 would converge quickly since we only add a small branch to the original Model1.

3.2. Pose Tracking

Table 2 demonstrates the performance of our SOPT tracker on PoseTrack partial test set. Compared with the result reported in [2], our SOPT algorithm achieves better performance in both pose estimation and tracking performance. Since the SOPT tracker does not exploit the multi-frame information and operate on a simple association mode, the SOPT tracker can runs realtime pose tracking (at ∼25 FPS) on a TITAN X Pascal GPU.

Joint	Model1	Model2
Head	40.3	73.9
Shoulder	70.1	72.6
Elbow	65.1	66.1
Wrist	54.1	54.5
Hip	59.5	60.4
Knee	54.9	55.3
Ankle	47.9	48.1
Total	54.8	62.4

Table 1. AP comparison on PoseTrack validation set.

Joint	AP (%)	MOTA(%)
Head	73.6	60.6
Shoulder	70.7	58.5
Elbow	62.8	36.7
Wrist	53.5	29.6
Hip	57.0	33.5
Knee	54.3	36.0
Ankle	48.4	28.8
Total	60.9	41.9

Table 2. SOPT Results on PoseTrack partial test set.

References

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1

[2] U. Iqbal, A. Milan, and J. Gall. Posetrack: Joint multi-person pose estimation and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1