Temporal Keypoint Matching and Refinement Network for Pose Estimation and Tracking

Chunluan Zhou Zhou Ren Gang Hua

Wormpex AI Research
czhou002@e.ntu.edu.sg, {renzhou200622, ganghua}@gmail.com

1 Influence of temporal window size

We use a temporal window $[t-\tau, t+\tau]$ with $\tau = 5$ for keypoint matching and refinement in all experiments in the paper. Here, we analyze the influence of the temporal window size on the performance of human pose estimation and tracking. We conduct this analysis on the PoseTrack 2017 dataset [1]. For this analysis, we use ResNet-101 and ResNet-152 as detection and pose estimation backbones respectively. Table 1 shows the results with different values of τ. $\tau = 0$ indicates keypoint refinement is not applied. From Table 1, we can see that temporal keypoint refinement can improve the performance for both pose estimation and tracking ($\tau > 0$ vs $\tau = 0$). The performance improves with the increase of τ. After $\tau > 5$, further increasing τ does not help much.

2 Network structure

The deep convolutional neural network used in our approach consists of a network backbone, a keypoint prediction module, a keypoint matching module and a keypoint refinement module as shown in Fig. 1. We use either Resnet-152 [2] or HRNet [3] as the network backbone. The structures of the three modules are given in Table 2. Each module comprises three basic blocks followed by several layers for producing the output. The keypoint prediction and refinement modules have the same structure.

References

1. Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L.: Posetrack: A benchmark for human pose estimation and tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
3. Ke, S., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

	$\tau = 0$	$\tau = 1$	$\tau = 3$	$\tau = 5$	$\tau = 7$
mAP	76.1	76.8	77.7	**78.2**	78.1
MOTA	67.3	68.1	69.2	69.6	**69.8**

Table 1. Influence of temporal window size.

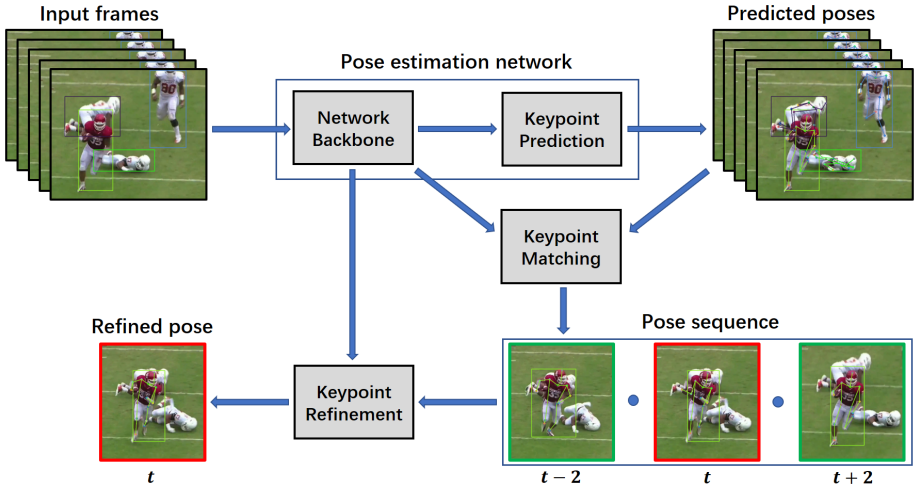

Fig. 1. Network structure.

Module	Basic block (×3)	Output
Prediction	Deconv $[\times 2, 256]$, BN, ReLU	Conv $[1 \times 1, 17]$
Matching	Conv $[3 \times 3, 256]$ (×3), Deconv $[\times 2, 256]$, BN, ReLU	FC [256] (×2), FC [2]
Refinement	Deconv $[\times 2, 256]$, BN, ReLU	Conv $[1 \times 1, 17]$

Table 2. Module structures. Deconv $[\times 2, 256]$ represents a deconvolution layer with upsampling scale of 2 and channel number of 256. Conv $[1 \times 1, 17]$ represents a convolution layer with kernel size of 1×1 and channel number of 17. FC [256] represents a fully connected layer with 256 output units. $(\times n)$ denotes repeating a layer n times. BN represents a batch normalization layer.