Hierarchical Temporal Transformer for 3D Hand Pose Estimation and Action Recognition from Egocentric RGB Videos

Yilin Wen[1], Hao Pan[2], Lei Yang[3], Jia Pan[1], Taku Komura[1], and Wenping Wang[4]

[1] The University of Hong Kong
[2] Microsoft Research Asia
[3] Centre for Garment Production Limited, Hong Kong
[4] Texas A&M University

Abstract. Understanding dynamic hand motions and actions from egocentric RGB videos is a fundamental yet challenging task due to self-occlusion and ambiguity. To address occlusion and ambiguity, we develop a transformer-based framework to exploit temporal information for robust estimation. Noticing the different temporal granularity of and the semantic correlation between hand pose estimation and action recognition, we build a network hierarchy with two cascaded transformer encoders, where the first one exploits the short-term temporal cue for hand pose estimation, and the latter aggregates per-frame pose and object information over a longer time span to recognize the action. Our approach achieves competitive results on the H2O and FPHA benchmark. Extensive ablation studies verify our design choices.

Keywords: hand pose estimation, action recognition, hierarchical transformer

1 Introduction

Perceiving dynamic interacting human hands from the egocentric RGB video is fundamental yet challenging, as there are frequent self-occlusions between hands and objects, as well as severe ambiguity of action types judged from individual frames (*e.g.* see Fig. 1 where the actions of *pour milk* and *place milk* can only be discerned at complete sequences).

Unified frameworks [11,12,16] have been proposed to simultaneously address both 3D hand pose estimation and action recognition, based on the critical observation that the temporal context of hand poses helps resolve action ambiguity, using models like LSTM, GCN or TCN. However, we note that temporal information can also benefit hand pose estimation: while hands are usually under partial occlusion and truncation especially in the egocentric view, they can be inferred more reliably from neighboring frames with different views by temporal motion continuity. Indeed, this idea has not been fully utilized yet by [11,12,16]: [11,12]

Fig. 1: Image sequences from H2O [11], with frequent occluded hand joints and ambiguous action type judged by individual frames.

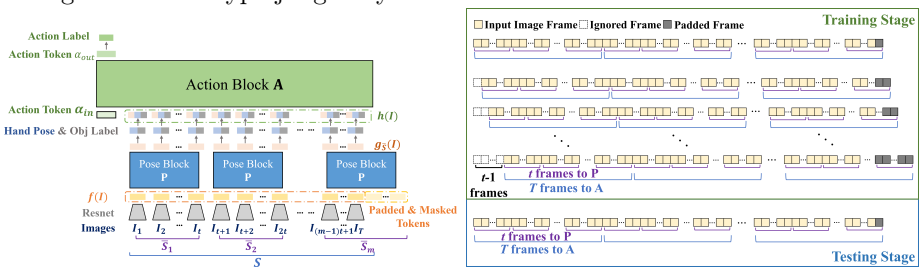


Fig. 2: Overview of our framework (Left) and the segmentation strategy for dividing a long video into inputs of **HTT** (Right).

perform image-based hand pose estimation, leaving the temporal dimension unexplored, and [16] jointly refines action and hand pose through hand-crafted multiple-order motion features and a complex iterative scheme.

We build a framework to exploit the temporal dimension for effective hand pose estimation and action recognition with a single feed-forward pass. To exploit the inter-frame relationship, we adopt the transformer architecture [13] which has demonstrated superior performance in sequence modeling. However, action and pose have different temporal granularity: while the action is related to longer time spans lasting for several seconds, the hand pose depicts instantaneous motions. Moreover, we notice that the action is usually defined in the form of "*verb* + *noun*" [7, 11], where *verb* can be derived from the hand motion and the *noun* is the object being manipulated. Therefore, we build a hierarchical temporal transformer with two cascaded blocks, to leverage different time spans for pose and action estimation, and model their semantic correlation by deriving the high-level action from the low-level hand motion and manipulated object label. Evaluation on H2O [11] and FPHA [7] verifies our competitive performances.

2 Methodology

Network design. Our network is visualized in Fig. 2, where given the egocentric RGB video S with T frames, we first feed each image to a ResNet-18 [9] feature extractor, and then pass the sequence of image features to our core network **HTT**, which is a hierarchical temporal transformer that outputs the per-frame 3D hand pose and action category for S respectively from two cascaded parts **P** and **A**. To implement the different time spans efficiently, videos are split into sub-sequences with a shifting window strategy.

The pose block **P** focuses on a narrower temporal receptive field with only t ($t < T$) consecutive frames, to improve the robustness under frequent invisible joints while also prevent confusion of local motion caused by overemphasis on temporally distant frames. Therefore, S is divided into consecutive segments $\texttt{seg}_t(S)$ by a shifting window strategy with window size t (see Fig. 2, left),

where tokens beyond the length T are padded but masked out from self-attention computation. **P** then processes each segment $\bar{S} \in \texttt{seg}_\texttt{t}(S)$ in parallel to capture the temporal cue for hand pose estimation, whose output per-frame token $g_{\bar{S}}(I)$ encodes for $I \in \bar{S}$ also the temporal cue from $\bar{S}$. We decode from $g_{\bar{S}}(I)$ the hand pose for I including the joint coordinates in the image plane P_I^{2D} and the joint depth to the camera P_I^{dep}. To supply the noun of an action [12], we also regress from $g_{\bar{S}}(I)$ the probability distribution O_I for the object.

The action block **A** leverages the full S to predict the action label, where we follow [2,3] to introduce an extra trainable token $\alpha_{\mathbf{in}}$ to aggregate the global information across S. The other T input tokens encode the per-frame information $h(I)$ of hand pose and object label, and also the image feature $g_{\bar{S}}(I)$ which may encode other useful cues like object appearance and hand-object contacts; we observe the best performance with such input data than alternative combinations. We classify the action for S from the first token α_{out} of the output sequence by **A**.

Implementation details. We set $T = 128$ and $t = 16$ as the respective maximum input sequence length for **A** and **P**, where T is derived from the limitation of available computational resources. To process a longer video, we split V into a clip set $\texttt{seg}_\texttt{T}(V)$ such that each clip $S \in \texttt{seg}_\texttt{T}(V)$ can be processed by **HTT**: V is first downsampled with a sampling ratio of 2 and further divided into consecutive clips by adopting the shifting window strategy with a window size T.

In training, while the ResNet-18 has its weight initialized from that trained on ImageNet, we do not pre-train our **HTT** on other datasets. We supervise the hand pose estimation by minimizing the $L1$-loss compared to the groundtruth, and the object and action classification with the standard cross-entropy loss. To augment sampling variations of training data, we offset the starting frame to each of the first t frames (Fig. 2, right), which ensure that both **P** and **A** consume different augmented data generated from the same sequence.

Testing stage computation. For a video V, we obtain the per-image 3D hand pose from **P** and the action category for V by voting from the output category among $S \in \texttt{seg}_\texttt{T}(V)$ (Fig. 2, right), therefore achieving efficient computation as each image is processed only once by both **P** and **A**.

3 Experiments

Result on H2O [11]. We report comparison with related works in the left part of Tab. 1, where we show competitive performance in both tasks. For hand pose estimation, our better performance compared with image-based methods [8, 11, 12] shows the benefits of using the temporal coherence for improved robustness.

Result on FPHA [7]. We report results in Fig. 5, where we outperform baseline methods for action recognition, and show competitive results for hand pose estimation against the video-based [4, 16] and image-based [12] methods.

Ablation Study. We verify the benefits of using short-term temporal cue for pose estimation in the upper-right of Tab. 1 and Fig. 3. Our $t = 16$ achieves the best performance, which shows enhanced robustness under invisible joints compared with $t = 1$, while avoids over-attending to distant frames and ensures sharp local motion compared with a long-term $t = 128$.

Table 1: Results of hand (*upper*) and action (*lower*) on the test split of H2O [11], for comparison with related works (*left*) and ablation study (*right*).

MEPE(in *mm*)	H+O [12]	LPC [8]	H2O [11]	Ours
L/R	41.42/38.86	39.56/41.87	41.45/37.21	**35.02/35.63**

MEPE(in *mm*)	$t=1, T=128$	$t=16, T=128$	$t=128, T=128$
L/R	40.12/40.62	**35.02/35.63**	36.41/39.36

	C2D [14]	I3D [1]	SlowFast [5]	H+O [12]	H2O [15] w/ ST-GCN	H2O [11] w/ TA-GCN	Ours
Acc.	70.66	75.21	77.69	68.88	73.86	79.25	**86.36**

	$t=16, T=16$ Cascaded **P**, **A**	$t=16, T=128$ Cascaded **P**, **A**	$t=16, T=128$ Parallel **P**, **A**
Acc.	74.38	**86.36**	80.17

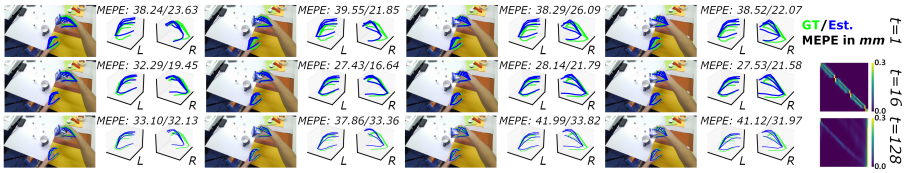
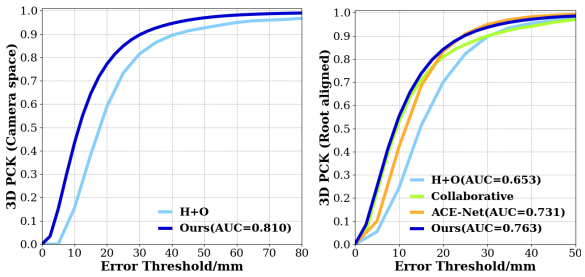


Fig. 3: Qualitative comparison of different t on H2O [11]. For $t = 16, 128$, the attention weights in the final layer of **P** is visualized.

Fig. 4: Visualization for weights of attention in the final layer of **A**, from the action token to the frames.



RGB-based methods	Acc.
Joule-color [10]	66.78
Two-stream [6]	75.30
H+O [12]	82.43
Collaborative [16]	85.22
Ours	**94.09**

Fig. 5: Results of hand (*left*) and action (*right*) on FPHA [7].

For action recognition, we verify the benefits of using a long time span and the cascaded design of **P** and **A** in the lower-right of Tab. 1, where the counterpart of parallel **P** and **A** are set by letting both **P** and **A** take the ResNet feature as the per-frame input token. For our design choice, we also visualize in Fig. 4 the attention weights of **A**, regarding a video of *take out espresso* whose action can only be judged by the last few frames depicting the process of taking the capsule out of the box, where correspondingly these key frames receive most attentions.

4 Conclusion

We have proposed a unified framework for 3D hand pose estimation and action recognition from an egocentric RGB video, to cope with the challenge of self-occlusions and action ambiguity. Our core network is a hierarchical temporal transformer consisting of two cascaded parts, for modeling the semantic correlation between the two tasks and leveraging different time spans according to their temporal granularity. Evaluations verify the effectiveness of our method.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. Association for Computational Linguistics (2019)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (2021)
4. Fan, Z., Liu, J., Wang, Y.: Adaptive computationally efficient network for monocular 3d hand pose estimation. In: European Conference on Computer Vision. pp. 127–144. Springer (2020)
5. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
6. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1933–1941 (2016)
7. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 409–419 (2018)
8. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 571–580 (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Hu, J.F., Zheng, W.S., Lai, J., Zhang, J.: Jointly learning heterogeneous features for rgb-d activity recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5344–5352 (2015)
11. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10138–10148 (2021)
12. Tekin, B., Bogo, F., Pollefeys, M.: H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4511–4520 (2019)
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
14. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
15. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)

16. Yang, S., Liu, J., Lu, S., Er, M.H., Kot, A.C.: Collaborative learning of gesture recognition and 3d hand pose estimation with multi-order feature analysis. In: European Conference on Computer Vision. pp. 769–786. Springer (2020)