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

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

[1]The University of Hong Kong [2]Microsoft Research Asia [3]TransGP [4]Texas A&M University

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Task

Unified framework for both 3D hand pose estimation and action recognition from the egocentric RGB video.

Key Designs

- ***Leverage the temporal information for both pose and action,*** to cope with:
  - frequent self-occlusions between hands and objects.
  - severe ambiguity of action types judged from individual frames.
- ***Build a hierarchical temporal transformer with two cascaded blocks,*** to:
  - ✓ ***leverage different time spans*** for pose and action estimation.
  - ✓ ***model the semantic correlation*** by deriving the high-level action from the low-level hand motion and manipulated object label.

Framework

Hand Pose Estimation with Short-Term Temporal Cue

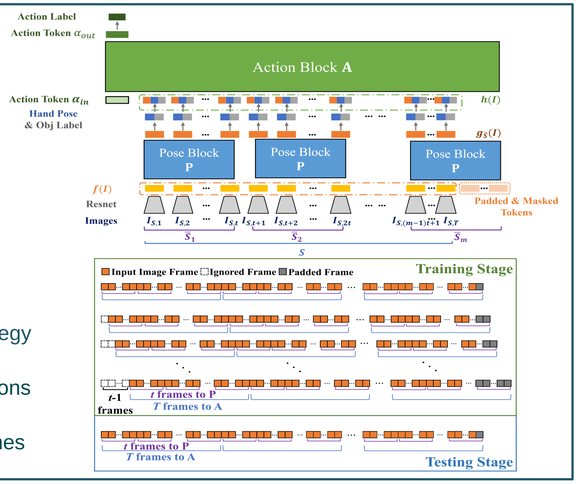
- Pose block $\boldsymbol{P}$ focuses on a narrower temporal receptive field to output per-frame 3D hand pose and manipulated object label.
- Input video is divided into consecutive segments by a shifting window strategy with window size t, segments are processed by $\boldsymbol{P}$ in parallel.

Action Recognition with Long-Term Temporal Cue

- Action block $\boldsymbol{A}$ uses the full video to predict the action label.
- The input of $\boldsymbol{A}$ leverages the per-frame predicted hand pose, object label and image feature.

Segmentation Strategy to Divide Long Videos into HTT Inputs

- Videos longer than T frames are divided into consecutive clips, by adopting the shifting window strategy with a window size T:
  - In testing stage, the hand pose are estimated by $\boldsymbol{P}$, the action category is voted from the predictions among segmented clips.
  - To augment training data, the starting frame for shifting window is offset to each of the first t frames



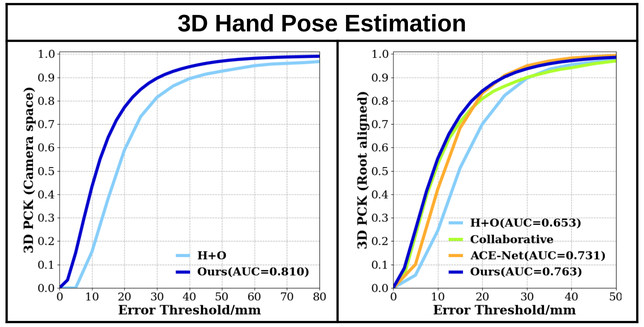
Result on H2O

Action Recognition							
	C2D	I3D	SlowFast	H+O	H2O w/ ST-GCN	H2O w/ TA-GCN	Ours
Acc.	70.66	75.21	77.69	68.88	73.86	79.25	**86.36**

3D Hand Pose Estimation (Camera Space)				
MEPE (in *mm*)	H+O	LPC	H2O	Ours
Left	41.42	39.56	41.45	**35.02**
Right	38.86	41.87	37.21	**35.63**

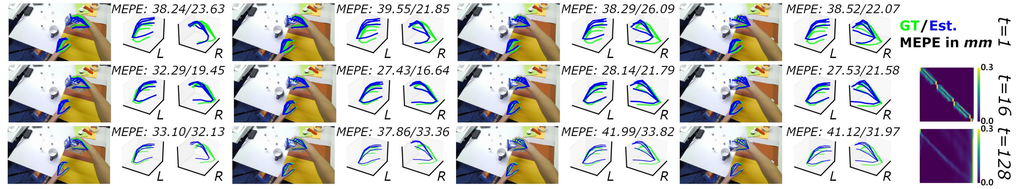
Result on FPHA

Action Recognition					
	Joule-color	Two-Stream	H+O	Collaborative	Ours
Acc.	66.78	75.30	82.43	85.22	**94.09**



†Work partially done during internships with Microsoft Research Asia.

Qualitative Results



Qualitative comparison of different t for hand pose estimation on H2O dataset.

The attention weights in the final layer of $\boldsymbol{P}$ is visualized for t=16,128.
Compared with t =1, our t =16 shows enhanced robustness.
Compared with t =128, our t =16 avoids over-attending to distant frames, therefore ensuring sharp local motion.

Visualization for weights of attention in the final layer of $\boldsymbol{A}$, from the action token to the frames.

Presented is a video of *take out espresso* from H2O dataset, whose down-sampled image sequence is shown in the top row.
The last few frames are the key for recognizing the action; in response our network pays most attention to these frames.