Supplementary Material: Functional Hand Type Prior for 3D Hand Pose Estimation and Action Recognition from Egocentric View Monocular Videos

Wonseok Roh[1]
paulroh@korea.ac.kr
Seung Hyun Lee[1]
easter3163@korea.ac.kr
Won Jeong Ryoo[1]
petac@korea.ac.kr
Jakyung Lee[1]
2023020917@korea.ac.kr
Gyeongrok Oh[1]
dhrudfhr98@korea.ac.kr
Sooyeon Hwang[1]
hsy506@korea.ac.kr
Hyung-gun Chi[2]
chi45@purdue.edu
Sangpil Kim[1*]
spk7@korea.ac.kr

[1] Department of Artificial Intelligence,
Korea University,
Seoul, Republic of Korea

[2] Electrical and Computer Engineering,
Purdue University,
West Lafayette, Indiana, USA

Overview

In this supplementary material, we supply further explanations and visualizations of our main paper, **"Functional Hand Type Prior for 3D Hand Pose Estimation and Action Recognition from Egocentric View Monocular Videos"**. We first provide additional experimental results to validate our proposed framework (Section A). We also describe implementation details for training (Section B). Next, we explain elaborate descriptions of hand type annotation process (Section C). We further provide more details for large-scale datasets (FPHA [1] and H2O [2]) and analysis of hand type distributions based on functional hand type taxonomy (Section D). Moreover, we supply extra details on the following project website: `https://kuai-lab.github.io/bmvc2023fhtp`.

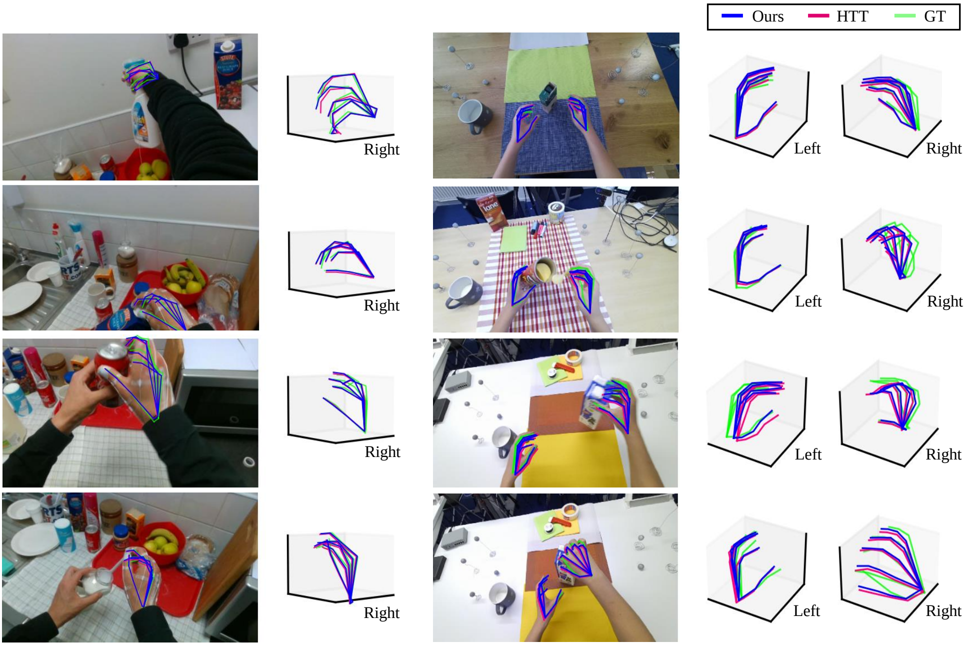
A. Additional Experimental Results

Qualitative Results In this section, we provide additional qualitative examples and analysis of our proposed model. In Fig. 1, we show various visualized results of the 3D hand poses

---

*Corresponding Author.

predicted by ours (see blue lines) and HTT [3] (see magenta lines) in 3D space and their projection to corresponding 2D image frames, compared with ground truth (see green lines). We observe that for both FPHA [1] (a) and H2O [2] (b), the gap between ours and ground truth is smaller than that between HTT and ground truth. These visualizations qualitatively verify that our model outperforms the baseline model by using functional hand type as semantic prior. In other words, utilizing hand type assists estimating hand pose in 3D space.



(a) FPHA Dataset (b) H2O Dataset

Figure 1: Qualitative comparison of 3D hand pose estimation results between our model (blue lines) and HTT (magenta lines) on the (a) FPHA [1] and (b) H2O [2] dataset. Both ground truth (green lines) and estimated hand poses are visualized in 3D space and projected to the corresponding 2D image frames. Note that the H2O provides labels for both hands, while the FPHA only includes information for the right hand.

B. Implementation Details

For training, we use two RTX 3090 GPUs with a batch size of 4, allowing for efficient processing and improved performance. Since the model converged at around 45 epochs, we trained it for a total of 45 epochs with a learning rate 3×10^{-5}, which is decreased 10 times every 15 epochs. We applied the Adam optimizer to train the dataset. Using PyTorch, we implement our experimental setup with the following settings. Our model architecture adopts K RGB frames, resized to 480×270, as the input sequence. All K input sequence images are sliced into k segments and fed into ResNet-18 pre-trained on ImageNet to extract image features. Then transformer block takes these features and outputs temporal-dependent features. We set the dimension size of all features and tokens to 512. Moreover, to optimize our model efficiently, we apply data augmentation strategies following the previous method [3].

C. Annotation Details

Beyond traditional hand grasp types, we categorize hand types based on functionality. In addition, we newly define additional hand types to describe more real-life hand activities and utilize deep semantic knowledge of hand types. For example, classic taxonomy includes the *"Index Finger"* category for holding an object while extending the index finger. However, as shown in Fig. 2, we add two more categories for extending the index finger; *"Poking"* which indicates to poking into an object without deforming it, and *"Extended Index Curl"*, which represents a hand type that utilizes the index finger to deform an object by applying pressure on its tip. Here, we explain the details of our annotation process based on these function-based taxonomy.

We manually annotate hand type all frames of landmark datasets, considering both the appearance of the hand and the various functions the hand performs. To make consistent annotation, we inspect which hand type certainly appears in specific actions before annotating each frame individually. Explanation in more detail; for the *"Open Juice Bottle"* action of the FPHA [1] dataset (left side of Fig. 1 in the main paper), we label *"Tripod"* as holding the juice lid and *"Dynamic Tripod"* as opening the lid. Therefore, the frames just holding the lid are marked as *"Tripod"* and from the frames that start to open, they are marked as *"Dynamic Tripod"* since the hand type changes to manipulate operation. More specifically, when the thumb finger is out while opening the lid, the hand type converts to *"Thumb Up"*. If the hand is out of frames or occluded by the object, it is annotated as *"Invisible"*. Also, for the challenging and unclear hand type, we go over whole videos implementing the same action by other subjects and then annotate as the most closely matched hand type.

As shown in Fig. 3, we show continuous annotation examples where hand types change over time in the video sequence, following the annotation details above. In Fig. 3 (a), we provide frames of the action called *"Place Cappuccino"* that interact with the object (cappuccino box). While the left hand keeps the hand on the desk (*"Relaxed Hand"*), the type of the right hand changes as it interacts with the cappuccino box. The right hand type starts as *"Large Diameter"* while holding the box and converts into a *"Fist"* through the intermediate stage of *"Relaxed Hand"* after putting the box down. Further, in Fig. 3 (b), we visualize frames of the action called *"High Five"* that interact with the other's hand. Even though the shape of the hand is the same throughout high-fiving, hand type differs as *"Open Hand"* or *"Dynamic Flatten"* based on whether it interacts with the other person's hand.

D. Analysis of Datasets and Hand Type Distributions

We train and evaluate overall performance on two landmark datasets (FPHA [1] and H2O [2]).

FPHA The FPHA [1] (First-Person Hand Action) dataset records 45 action categories from 6 subjects with an Intel RealSence SR300 RGB-D camera mounted on the subject's shoulder. It designs various action scenarios interacting with 26 different objects. Only J=21 joints of the subject's right hand are annotated using 3D poses from the wearable magnetic sensors. It contains 1,175 sequences and separates into 1:1 settings with 600 and 575 videos for the training and evaluation phase in each. Both steps include all subjects and actions.

H2O The H2O [2] (2 Hands and Objects) dataset captures 4 subjects performing 36 actions, interacting with 8 3D objects. Unlike the FPHA [1] dataset, which only provides labels for the right hand, H2O provides 3D labels for both hands with the total number of $J = 21 \times 2$ joints. It includes 569 videos for the training, including all actions for the first 3 subjects, 122 videos for the validation, and 242 videos of the remaining subject unseen in the training for the test.

Hand Type Distribution In this paper, we introduce a newly defined taxonomy of 31 hand types based on functionality. We visualize the distribution of our proposed hand type taxonomy with a radial diagram in Fig. 4. We first categorize the mainly used hand types into *Isolated* and *Interaction* based on whether the hands interact with objects (or other people). We then classify the *Interaction* types into *Grasp* and *Control*, reflecting the role of the hand in the scene. These categories are further subdivided according to the appearance of the hand and the degree of object manipulation. Finally, function-based taxonomy that comprehensively encompasses real-life hand movements allows explaining various hand action scenarios. With our new taxonomy, we annotate hand types across all frames of two datasets.

To investigate the hand type distributions of both FPHA [1] (Fig. 5 (a)) and H2O [2] (Fig. 5 (b)) datasets, we illustrate them via histograms and scatter plots. We first explore the number of frames per hand type with the histograms (Fig. 5 left) to observe how frequently each hand type occurs within whole hand video sequences. The x-axis represents the hand type, and the y-axis shows the number of frames. As shown in Fig. 5 (left), the distributions of the two datasets exhibit different patterns. In the case of the FPHA, 29 of the 31 hand types are identified, with *"Large Diameter"*, *"Disk Grip"* and *"Thumb Tucked"* being the most common. For the H2O, 24 of the 31 hand types are recognized, with *"Large Diameter"*, *"Relaxed Hand"* and *"Parallel Extension"* being the most prevalent. We find that the FPHA operates more diverse hand types than H2O for more action scenarios.

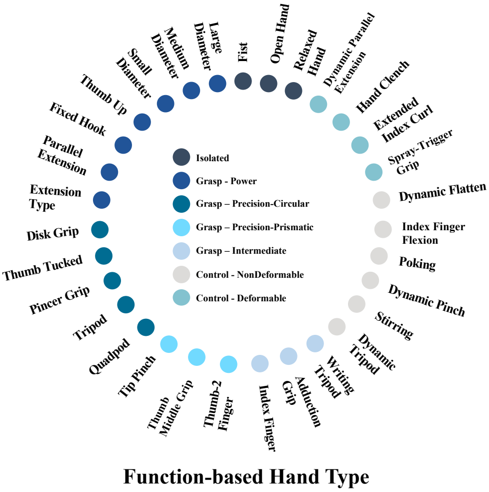
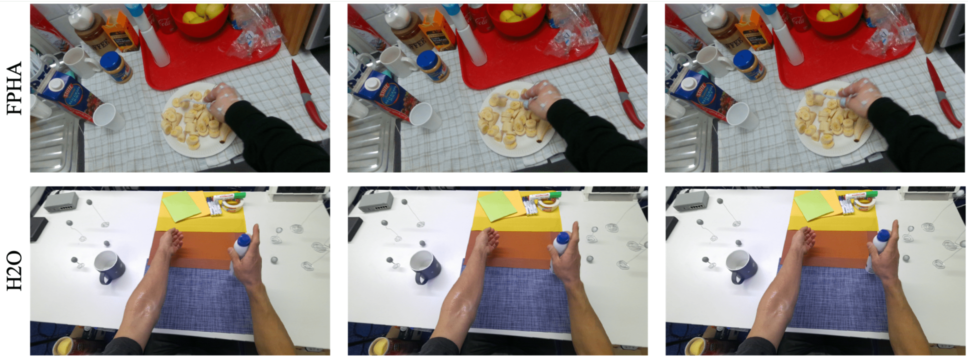
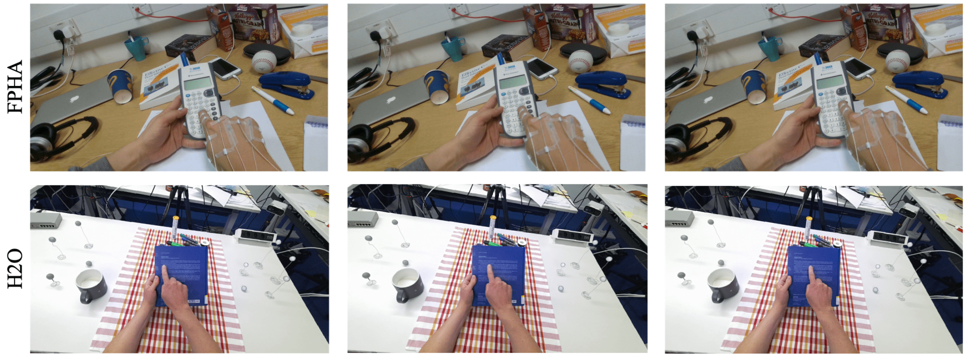


Figure 4: Radial diagram of function-based hand type taxonomy. To explain more real-life hand behaviors, we carefully design a taxonomy based on the functioning perspective.

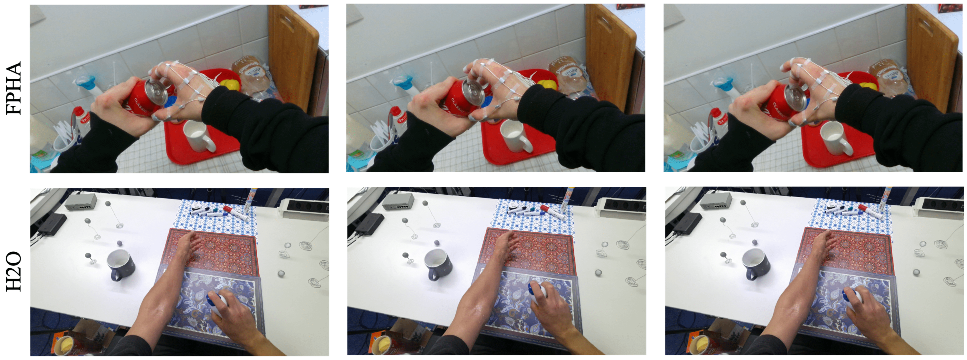
Also, we visualize the distribution of hand types across each action label with scatter plots (Fig. 5 right); the x-axis indicates the action of each dataset, and the y-axis represents the hand type. The circle size portrays the appearance frequency of each hand type within each action video clip. As shown in Fig. 5 (right), the FPHA presents monotonous distributions of hand types per action, while the H2O shows a relatively balanced distribution. This difference is because, unlike the H2O, which focuses on both hands, the FPHA annotates only on the right hand and records simple actions repeatedly. Ultimately, FPHA utilizes 29 hand types for 45 action scenarios but is monotonous, and H2O uses 24 hand types for 36 action scenarios but is relatively harmoniously distributed. Note that we further visualize hand type distributions of each left and right hand of the H2O dataset in Fig. 6.



(a) Index Finger

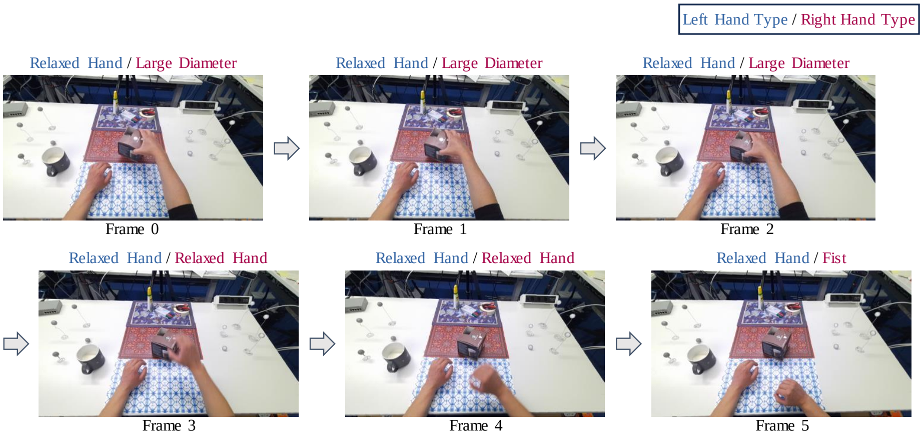


(b) Poking

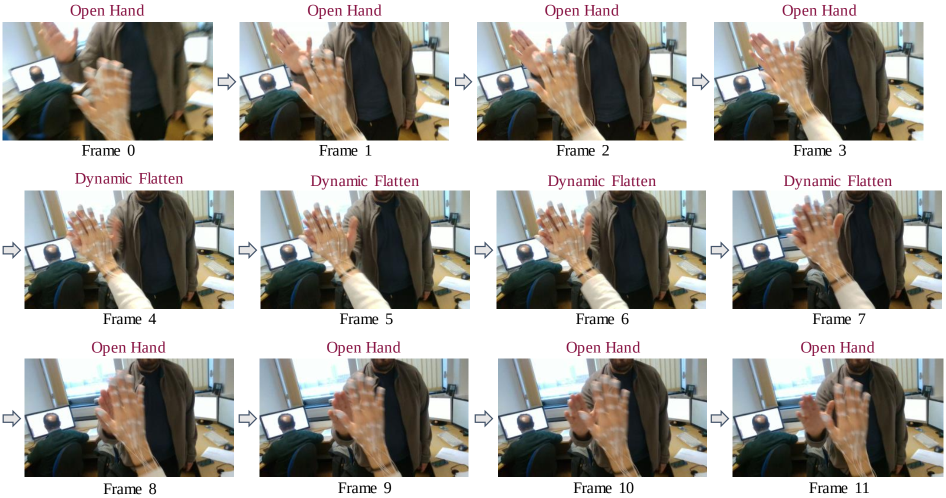


(c) Extended Index Curl

Figure 2: Examples of action scenes in which the variation of hand pose is similar, but the action labels are different according to the temporal context of hand function and object types. (a) *"Index Finger"*, (b) *"Poking"*, and (c) *"Extended Index Curl"* are all hand types that use index fingers, but their functions vary depending on the action and object they interact with. *"Index Finger"* represents holding an object while extending the index finger, *"Poking"* represents poking into an object without deforming it, and *"Extended Index Curl"* represents utilizing the index finger to deform an object by applying pressure on its tip.

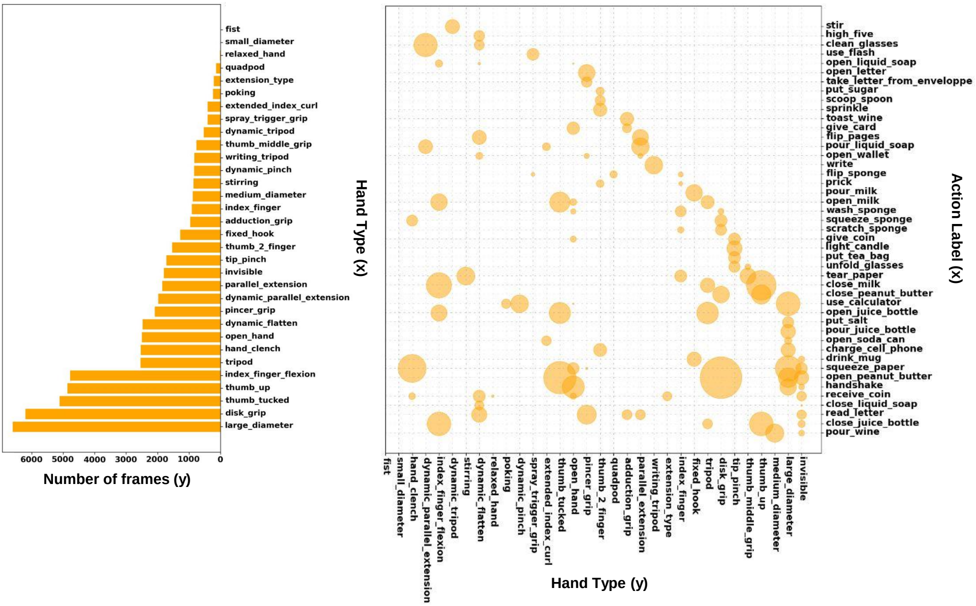

Left Hand Type / Right Hand Type
Relaxed Hand / Large Diameter
Frame 0
Relaxed Hand / Large Diameter
Frame 1
Relaxed Hand / Large Diameter
Frame 2
Relaxed Hand / Relaxed Hand
Frame 3
Relaxed Hand / Relaxed Hand
Frame 4
Relaxed Hand / Fist
Frame 5

(a) H2O Dataset

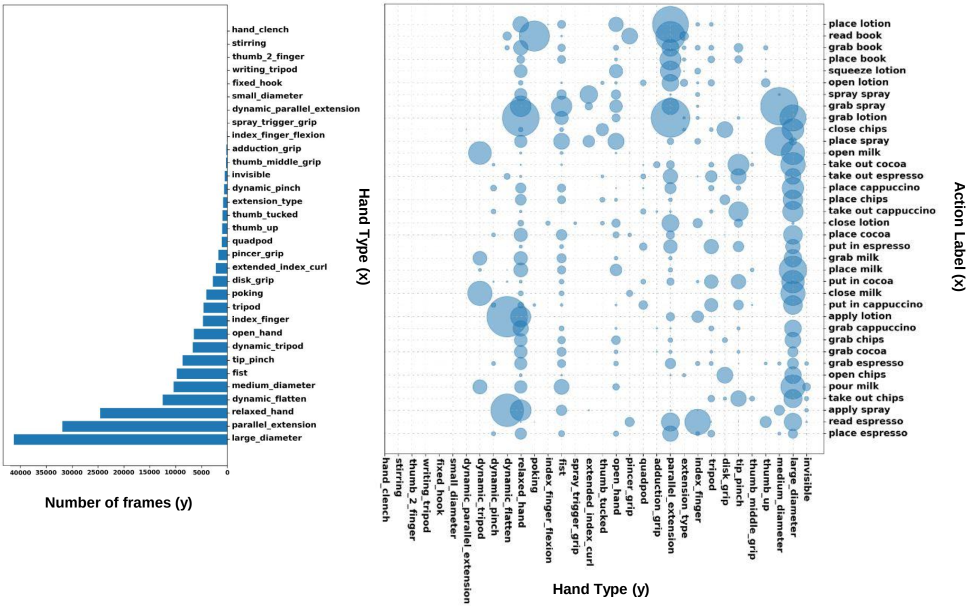


(b) FPHA Dataset

Figure 3: Examples of continuous video frames with hand type annotation. The hand type of each frame changes over time in the video sequence. In (a) and (b), we show frames of the *"Place Cappuccino"* action (H2O [2]) in which the hand interacts with the box and the *"High Five"* action (FPHA [1]) in which the hand interacts with the other's hand, respectively.

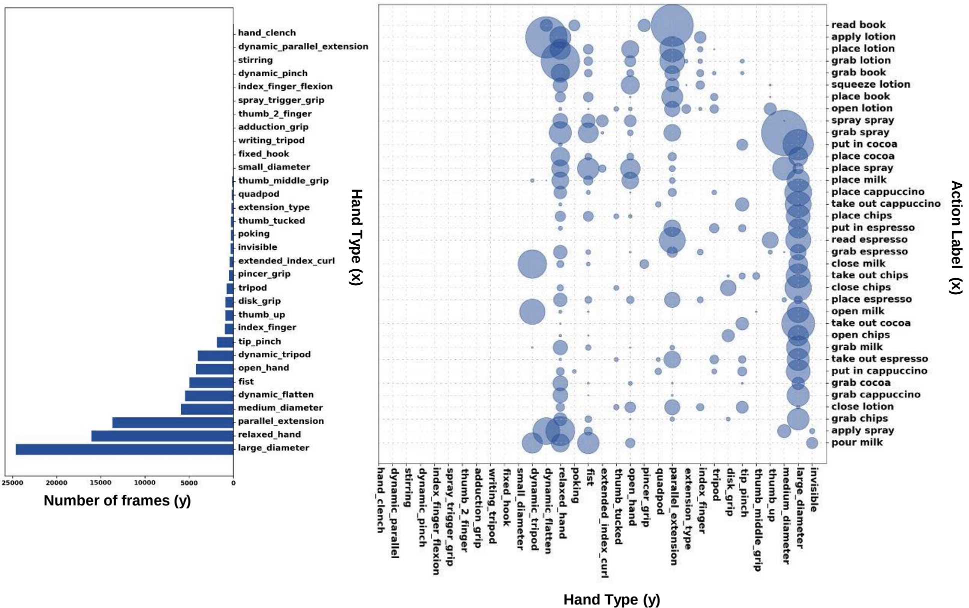


(a) FPHA Dataset

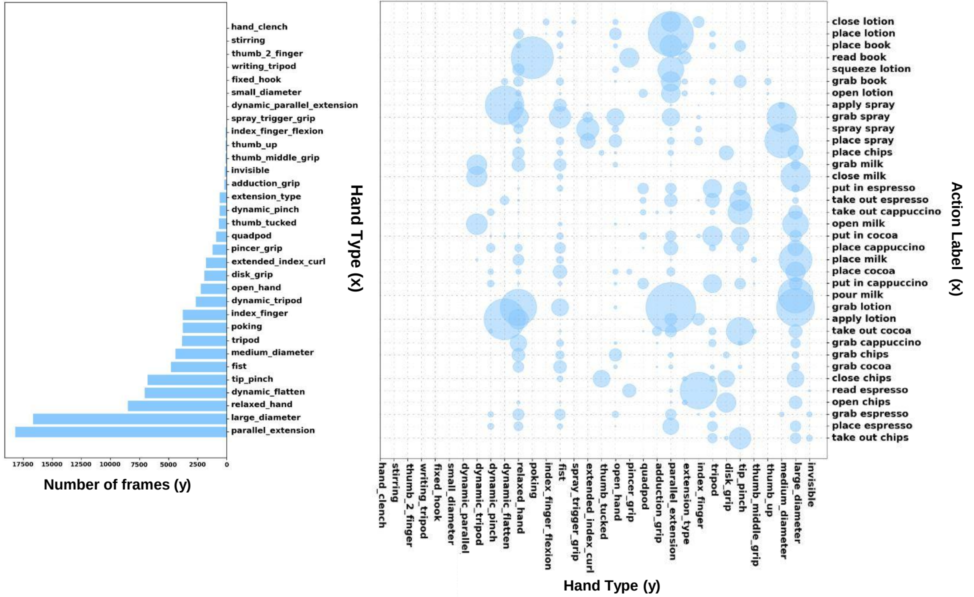


(b) H2O Dataset

Figure 5: Hand Type Distributions of the (a) FPHA [1] Dataset (Orange) and (b) H2O [2] Dataset (Blue). The histograms (left) show the number of frames per hand type; the x-axis represents the hand type, and the y-axis indicates the number of frames. The scatter plots (right) present the distribution of hand types across each action label; the x-axis represents the action of each dataset, and the y-axis indicates the hand type. The size of the circles illustrates how often each hand type appears within the hand action video sequences.



(a) H2O Dataset (Left)



(b) H2O Dataset (Right)

Figure 6: Hand Type Distributions for the (a) Left hand and (b) Right hand of the H2O [2] Dataset. The histograms (left) show the number of frames per hand type; the x-axis represents the hand type, and the y-axis indicates the number of frames. The scatter plots (right) present the distribution of hand types across each action label; the x-axis represents the action of each dataset, and the y-axis indicates the hand type. The size of the circles illustrates how often each hand type appears within the hand action video sequences.

References

[1] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018.

[2] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021.

[3] Yilin Wen, Hao Pan, Lei Yang, Jia Pan, Taku Komura, and Wenping Wang. Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric rgb videos. In *CVPR*, 2023.