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

Abstract

Current methods for egocentric view action recognition often face challenges in perceiving dynamic hand movements relying solely on geometrical or physical information. In this work, we effectively address this problem by gaining insights into the correlation between functional hand configurations and objects, which improves the detailed interpretation of real-world scenarios. To this end, we introduce a practical taxonomy of hand types based on the functioning perspective and utilize it for per-frame hand type labeling on existing datasets. We also propose a novel hand action recognition framework considering semantic details of the hand type as prior. This approach boosts the network's understanding of the continuous hand interaction throughout the action sequence. Our whole pipeline consists of three main modules: (1) Feature Extraction, (2) Egocentric Knowledge Module, which estimates 3D hand pose, object category, and hand type leveraging short-term cues, and (2) Egocentric Action Module, which aggregates per-frame knowledge, including text embeddings of hand type, over a longer time. In our extensive experiments with large-scale benchmarks, FPHA and H2O, our model outperforms current state-of-the-art methods, demonstrating its superior performance.

*Corresponding Author.

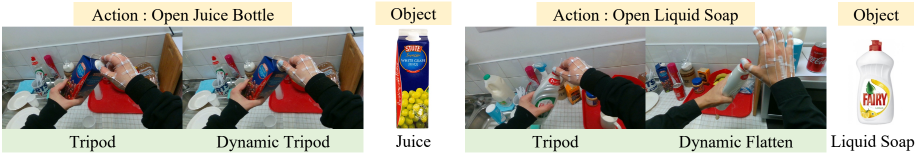

Figure 1: Examples of actions and corresponding hand types from the FPHA dataset [15]. While each action shares the goal of opening objects, the specific interactions between hands and each object are different. Further, both activities involve holding and opening the lid, but the semantic functions of these motions differ significantly. To highlight the continuously repeated shape during the operation, we represent it using *"Dynamic"*.

1 Introduction

Understanding dynamic interacting human hands is a promising computer vision task because people use their hands to handle objects and communicate with others throughout daily activities. Recently, various hand-based visual applications have been proposed, such as human-robot collaboration [13, 14, 33, 47] and imitation learning [31, 36, 37]. Moreover, the advancement of low-cost wearable sensors and VR/AR technologies motivates the computer vision community to tackle egocentric view hand action recognition.

In real-world egocentric view scenarios, substantial occlusion and truncation often occur, especially when the hand actively interacts with other hands or objects. To address these issues, recent approaches [8, 24, 26, 28, 40, 43, 44, 49, 51, 52] in hand action recognition primarily focus on the temporal context of 3D hand position or high-level object labels. However, these methods still face challenges in perceiving dynamic hand movements relying solely on geometrical information. In this work, we go beyond mere physical information and focus on the semantical hand interactions to provide valuable details for a comprehensive understanding of hand action sequences. For example, as illustrated in Fig. 1, there are semantic differences between everyday hand actions such as *"Open Juice Bottle"* and *"Open Liquid Soap"*. Although each action shares the goal of opening objects, the specific interactions between hands and each object differ. Therefore, gaining insights into the semantic relationships between functional hand configurations and other hands or various objects is essential for addressing complex real-world hand scenarios. From this observation, we consider utilizing contextual knowledge of hand types to understand complex hand interactions above simply relying on the physical hand positions.

One of the best ways to specify hand type at semantic level is hand type taxonomy. The concept of hand type represents a figurative expression of human intention when conducting tasks with hands. Most previous studies [2, 6, 11, 12, 34, 39] organize hand type taxonomies based on the properties of objects since humans usually use the similar grasp types for specific objects. In other words, they design and utilize hierarchical hand type categorization criteria based on object grasp manner. Although this hierarchy helps explain the appearance of the hand, this approach overlooks the importance of hand functionality or how the hand operates in various tasks. The function of the hand plays a crucial role in effectively understanding each action, as it accurately reflects the hand-object correlation as well as the intention behind the movement. To this end, we carefully redefine a practical taxonomy of hand types focused on the functioning perspective for diverse hand-related vision tasks, including egocentric view action recognition. Then, we supervise the hand action recognition network with hand type labels annotated with newly defined taxonomy.

Built upon [49] framework, which considers the high-level action recognition task as a mixture of two low-level tasks: 3D hand pose estimation and object classification, we especially utilize hand type estimation network to provide valuable semantic cues. Our model consists of three main modules: (1) Feature Extraction, (2) Egocentric Knowledge Module, which estimates 3D hand pose, object category, and hand type leveraging short-term cues, and (3) Egocentric Action Module, which aggregates per-frame knowledge, including text embeddings of hand type, with long-term cues. To the best of our knowledge, we are the first to adopt the usefulness of semantic cues of hand type for hand action recognition. Our proposed framework outperforms the existing state-of-the-art works on 3D hand pose estimation and action recognition. To summarize, our main contributions are listed as follows:

- We newly define the practical taxonomy of functional hand types based on the functioning perspective for diverse hand-related vision tasks, including hand action recognition. We also provide precise hand type annotations for existing datasets.
- We present a novel hand action recognition framework utilizing semantic details of the hand type in each frame. This approach guides the entire network to learn deep understandings of the continuous interaction across hands and objects throughout the action sequence. To the best of our knowledge, our work is the first to leverage the semantic knowledge of hand type in the context of hand perception.
- We analyze the effectiveness of our proposed hand type prior framework on large-scale benchmarks, including FPHA and H2O. Extensive experiments validate that our method generates a new State-of-the-Art score on 3D hand pose estimation and action recognition from egocentric view monocular videos.

2 Related Work

3D Hand Pose Estimation & Hand Action Recognition Hand action typically involves the interaction of hands and objects. 3D hand position provides important information about the object geometry and grasp type, which is positively correlated with hand action [3, 4, 41, 53]. Therefore, the hand pose is an influential feature for hand action recognition [15, 44]. Recently, CNN-based approaches [35, 46], and graph convolutional network-based methods [24, 40, 51] have learned important hand pose information from extracted meaningful spatial features. However, since hand actions are not static representations, it is difficult to perceive the continuous movement of hands. Thus, it is crucial to consider temporal information to improve understanding of hand activities. Alternative action recognition models such as temporal CNN-based [20, 22, 30, 52], LSTM-based [27, 28, 44], and two-stream networks [5, 9, 10, 42] appear, yet they still rely on either information. To consider both spatial and temporal information, current state-of-the-art works [25, 29] introduce transformer-based approaches with the multi-head self-attention network that helps find the relationship between the input sequences. From these observations, we adopt hierarchical transformers model architecture to learn not only spatial information through positional encoding, but also geometric 3D hand position with short-term temporal cues and semantic action flow with long-term temporal cues hierarchically at once.

Hand Type Taxonomy The relation between hand grasp type and object has been widely studied for decades [3, 16, 23]. The grasp types reveal the characteristics of the object because people generally use the same or similar hand grasp types for particular things. Early work by Schelesinger *et al.* [39] first categorized hand grasp type into six major types based

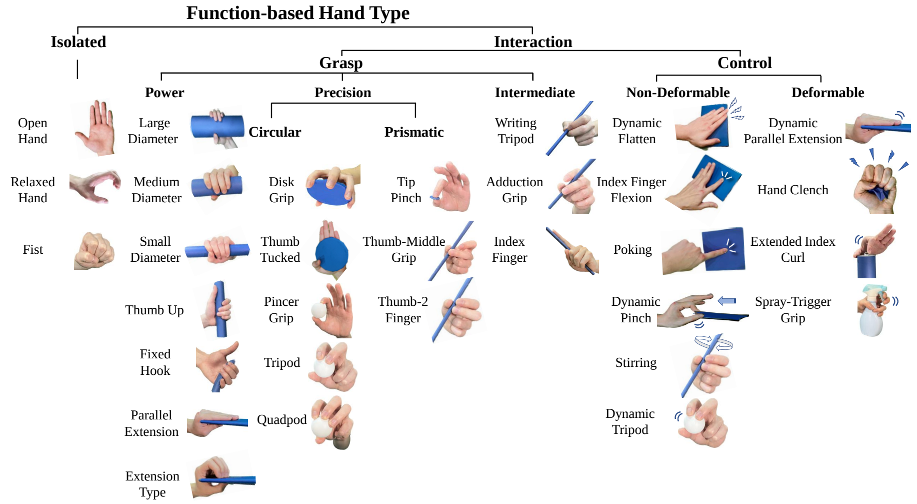

Figure 2: Taxonomy of hand types based on functionality. We categorize the mainly used hand types focusing on the role of the hand. Beyond static hand grasp types, we further define dynamic hand types to *Control* to explain more real-life hand behaviors.

on object shape, hand surfaces, and hand shape. Also, Napier *et al.* [34] divided hand movements into two main groups, prehensile and non-prehensile, and established the concept of power and precision grasp types. Furthermore, in recent hand-related works [19, 32, 50], the grasp taxonomy proposed by Feix *et al.* [12] has been widely used in hand analysis. Understanding the hand grasp type helps the model recognize the user's intention more accurately and improves the accuracy of hand action recognition. Therefore, we precisely redefine the taxonomy of hand grasp types based on functionality for practical action comprehension.

3 Function-Based Hand Type Taxonomy

Insight into the context between functional hand configurations and objects could significantly advance understanding of hand action. In this work, we consider the semantic prior of the human-level hand type as a critical indicator of perceiving egocentric view scenes. To achieve this goal, we first carefully design a useful taxonomy of hand types based on the functioning perspective and make use of them for effective dataset curation.

Redesign Taxonomy The grasp type is a figurative expression of human intention when performing tasks with hands, and it is closely related to hand action. Furthermore, since we interact with the other person's hand or object to perform many activities, it is very important to identify the hand type in understanding the relationship between the hand and the others. Most previous studies [1, 3, 18] have defined hand type taxonomies based on the characteristics of objects interacting with the hand because people usually use the same or similar grasp types for specific objects. They design and utilize hierarchical hand type categorization criteria based on grasp manner. We empirically observe that this hierarchy helps understand the appearance of the hand. However, to better understand the activity in the egocentric view video clip, it is more beneficial to categorize based on the function and role of the hand as well as the appearance of the hand. In other words, there are limitations to traditional taxonomy in explaining various hand action scenarios in everyday life. Our carefully designed taxonomy starts from this observation.

As shown in Fig. 2, we first categorize the mainly used hand types into two groups: *Isolated* and *Interaction*. *Isolated* represents independent hand types that do not interact with other people or objects, and *Interaction* describes hand types that are actively interacting with other people or objects. We then organize the hand types as *Interaction* into *Grasp* and *Control*, focusing on the functions in the activity. *Grasp* includes hand types that hold an object even over time, which are further divided into *Power*, *Precision*, and *Intermediate* depending on the hand's appearance. We also distinguish between *Circular* and *Prismatic* pivoting on the shape of the interacting objects. On the other hand, *Control* includes hand types that manipulate objects over time. Here, we classify hand types, concentrating on the deformation of the object. For example, the hand type that turns the lid does not deform the object, but the hand type that crumples the paper or the hand type that opens the can deform the object. Thus, we differentiate them as *Non-Deformable* and *Deformable* according to these criteria.

Dataset Annotation Ultimately, we perform hand type annotation for all frames of the FPHA [15] and H2O [24] dataset based on the newly defined taxonomy. We implement hand type labeling on the right hand for FPHA, as it contains information only for the right hand. However, for the H2O dataset, which provides data for both hands, we label both hands. Note that more details about dataset annotation are provided in supplemental material.

4 Learning Functional Hand Type

In this section, we advocate for leveraging human-level hand type knowledge to provide semantically rich cues beyond using simply physical-level hand pose information. In particular, we explicitly guide the action recognition pipeline via rich semantic prior of the hand type based on hand type taxonomy that we proposed in Sec. 3.

It is noteworthy that temporal information for estimated 3D hand position and high-level object labels enhances action recognition accuracy. Although existing techniques [24, 44, 49] benefit from the generous geometric potential of 3D position, they do not cover diverse real-world scenarios. Specifically, they overlook the semantic role of hand motions in various scenarios and still face challenges in perceiving dynamic hand movements relying solely on geometrical or physical information. Therefore, we propose to utilize semantic knowledge of hand type in each frame to encourage the network to understand the continuous interaction between the primary hand and the assistive hand/object in the video clip.

4.1 Overview

We illustrate the outline of our framework in Fig 3. Built upon hierarchical temporal transformer (HTT [49]), which considers the high-level action recognition task as a mixture of two low-level tasks: 3D hand pose estimation and object classification, we especially utilize hand type estimation network to provide valuable semantic cues. Our model consists of three main modules: (1) Feature Extraction, (2) Egocentric Knowledge Module, which estimates 3D hand pose, object category and hand type leveraging the short-term temporal cue and (3) Egocentric Action Module, which aggregates per-frame pose, object and hand type information over a longer time span. First, our model takes aligned 2D video clip $V = \{X_i \in \mathbb{R}^{3\times H\times W} | i = 1,...,K\}$ consisting of K frames as input, which are converted into feature vector F_I containing fine details. Then we employ temporal-dependent features F_H from the Local Transformer to estimate the per-frame 3D hand pose with feature vector

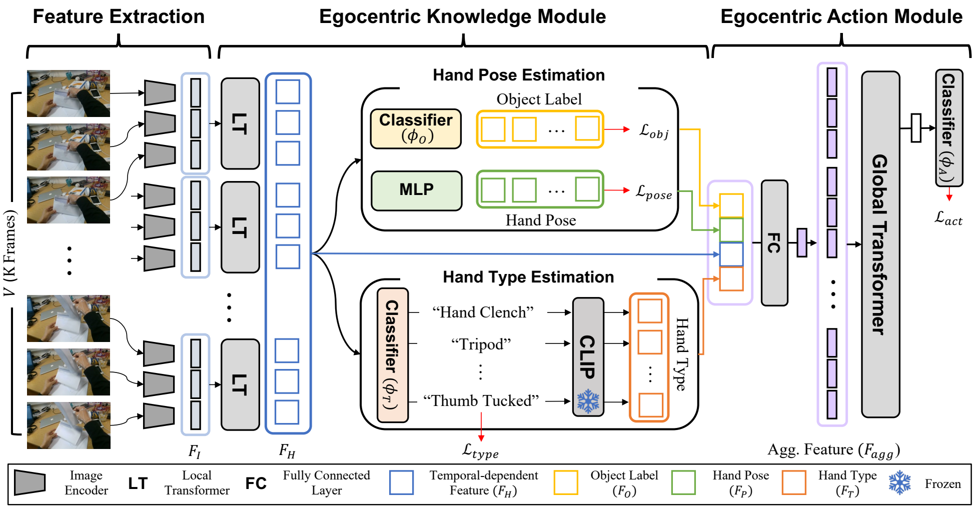

Figure 3: Overview of our proposed model. Our framework consists of three main modules: (1) Feature Extraction, (2) Egocentric Knowledge Module, which estimates 3D hand pose, object category and hand type leveraging short-term temporal cues, and (3) Egocentric Action Module, which aggregates per-frame pose, object and hand type information over a longer time span.

F_P, which contains geometric potential, and the category of interacting objects with feature vector F_O. Additionally, the hand type estimation network takes F_H and outputs hand type feature vectors F_T. Here, we utilize language models to provide the deep semantic cues of hand type based on our proposed functional hand type taxonomy. Subsequently, the Egocentric Action Module aggregates the predicted embeddings: hand position F_P, object category F_O, and hand type F_T for action recognition.

4.2 Egocentric Knowledge Module

To construct Egocentric Knowledge Module input sequence, we first divide the long video clip V into m consecutive segment $\texttt{seg}_k(V) = (\bar{V}_1, \bar{V}_2, ..., \bar{V}_m)$, where m denotes $\lceil K/k \rceil$. In order to capture the temporal cue of consecutive segment for hand pose estimation, the module processes each segment $\bar{V} \in \texttt{seg}_k(V)$ in parallel. Then transformer takes the sequence of per-frame feature vector F_I from the image encoder and outputs temporal-dependent features F_H. To decode the hand pose information, these features F_H are fed into simple MLP layers, yielding joint coordinates in the 2D image plane $P^{2D} \in \mathbb{R}^{J \times 2}$ and the joint depth to the camera $P^{dep} \in \mathbb{R}^{J \times 1}$. We train the 3D pose estimation module to minimize the following pose loss (L1-loss):

$$\mathcal{L}_{\text{pose}} = \frac{1}{J}(||P^{2D} - P^{2D}_{gt}||_1 + \lambda_{\text{pose}}||P^{dep} - P^{dep}_{gt}||_1) \tag{1}$$

where J denotes hand joints and λ_{pose} is a hyperparameter to balance the different intensities of the 2D and depth losses. The 3D positions of the hand joints in the camera space $P^{3D} \in \mathbb{R}^{J \times 3}$ for I can be inferred operating the camera intrinsics.

In addition to using 3D hand pose information, which gives precise geometric knowledge about egocentric view scenarios, we advocate for leveraging hand type information to guide the entire network to learn deep semantic understanding. In particular, the context details of hand type can serve as a practical semantic key representing hand-object relationships for identifying hand actions. Thus, we introduce a simple but novel hand-type classification network ϕ_T to predict the hand type t_i for $i = \{1, 2, \ldots, N_t\}$ from temporal-dependent features F_H. Given the ground truth hand type label t_{gt} from the dataset annotation process (see Sec. 3), the target probability is defined as a one-hot vector w_t. For training, we formulate the following cross-entropy loss to train the hand type classification network ϕ_T:

$$\mathcal{L}_{type} = -\mathbb{E}_{t,w_t \sim \mathbb{D}} \left[\sum_{r \in N_t} w_t[r] \log \phi_T(t)[r] \right] \tag{2}$$

where $\mathbb{D}$ is (input) data distribution. Also, we predict the object category o_i in each frame with the object classification network ϕ_O. Similar to hand type classifier, the classifier ϕ_O is supervised to minimize the cross entropy loss $\mathcal{L}_{obj}$.

4.3 Boosting the Action Recognition with Hand Type Prior

We introduce the strategies using semantic prior of the human-level hand type, potentially improving not only action recognition but also 3D hand pose estimation. Specifically, we explicitly pilot the Egocentric Action Module to learn rich semantic hand type variant information via utilizing the power of the prevalent language model, Contrastive Language-Image Pre-Training [38] (CLIP). Given the candidate of hand type from the hand type classification network ϕ_T, we map the type number t_i with the corresponding text descriptions (*e.g.*, *"Hand Clench"*, *"Tip Pinch"*). Next, the text label of hand type goes through the large-scale pre-trained language model, which outputs text embedding vector F_T. Importantly, we consider these text embedding vector of predicted hand type descriptions as critical indicators to deliver semantic knowledge to the action recognition pipeline.

4.4 Egocentric Action Module

We adopt the previous approaches [7, 21] presenting trainable tokens to aggregate the global information across the input video clip V. Each token encodes short-term temporal information such as temporal-dependent features, hand pose, object label, and hand type details. We design the fully connected layer for each cues, which outputs features of the same dimension, then concatenates these aligned features for input of the Global Transformer as follows:

$$F_{agg} = \mathtt{FC}[F_H \oplus F_P \oplus F_O \oplus F_T] \tag{3}$$

where $\oplus$ indicates channel-wise concatenation of feature vectors and $\mathtt{FC}[\cdot]$ reduces the features into d-dim to fit in the token dimension of Global Transformer. After mixing features that potentially contain geometric and semantic knowledge, we feed these feature vector F_{agg} to Global Transformer, which outputs action tokens. Here, we utilize an action classification head ϕ_A to recognize action label a_i for $i = \{1, 2, \ldots, N_a\}$ from action tokens. For supervision, we formulate the following cross-entropy loss to train the action classification head ϕ_A:

$$\mathcal{L}_{act} = -\mathbb{E}_{a,w_a \sim \mathbb{D}} \left[\sum_{r \in N_a} w_a[r] \log \phi_A(a)[r] \right] \tag{4}$$

where w_a denotes one hot-encoded action labels and $\mathbb{D}$ represents (input) data distribution.

4.5 Training

Our entire network is trained end-to-end by minimizing the following loss $\mathcal{L}_{\text{total}}$:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{act}}\mathcal{L}_{\text{act}} + \frac{1}{K}\sum_{\bar{V}\in\texttt{seg}_{\text{k}}(V)}\sum_{I\in\bar{V}}(\lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \lambda_{\text{type}}\mathcal{L}_{\text{type}}) \tag{5}$$

where λ_{act}, λ_{pose}, λ_{obj} and λ_{type} are the hyperparameters of respective loss terms.

5 Experiments

Table 1: Comparison of our novel hand action recognition framework and the state-of-the-art models on the FPHA [15] and H2O [24] dataset. We report the classification accuracy of methods based on RGB videos. Note that the H2O dataset provides additional testing split videos, unlike the FPHA dataset, which provides only training and validation split.

	Joule-color [17]	Two Stream [9]	H+O [44]	Collaborative [52]	HTT [49]	Trear [25]	Ours
Accuracy (↑)	66.78	75.30	82.43	85.22	94.09	94.96	**95.13**

(a) Action recognition accuracy (%) on FPHA.

	C2D [48]	I3D [45]	SlowFast [10]	H+O [44]	ST-GCN [51]	TA-GCN [24]	HTT [49]	Ours
Val Accuracy (↑)	76.10	85.15	86.00	80.49	83.47	86.78	90.16	**91.80**
Test Accuracy (↑)	70.66	75.21	77.69	68.88	73.86	79.25	86.36	**89.67**

(b) Action recognition accuracy (%) on H2O.

5.1 Experimental Setups

Dataset We train and evaluate overall performance on two landmark datasets for action recognition from egocentric views: FPHA [15] and H2O [24]. These two datasets are collected in various indoor settings and have a frame rate of 30 frames per second. Both datasets provide the ground truth labels for hand pose, action, and object category, and we utilize them for supervision and evaluation. In this work, we annotate the hand type of all frames in these two datasets based on the newly defined taxonomy and use them for training. Note that detailed descriptions of each dataset are provided in the supplemental material.

Evaluation metrics To evaluate hand action recognition, we follow the official evaluation protocol of the hand action recognition task. We report classification accuracy over the validation and test split by comparing each video's predicted and ground truth action categories. Also, we evaluate the 3D pose estimation performance in the camera space and the root-aligned (RA) space, which aligns the estimated wrist position with the ground truth for each frame. We report the Percentage of Correct Keypoints (PCK) for joints [54] against different error thresholds and the corresponding Area Under the Curve (AUC). We also utilize Mean End-Point Error (MEPE) metrics for hands [54] in the camera and root-aligned space following HTT [49]. We provide all implementation details in the supplemental material.

5.2 Comparison with the State-of-the-Arts

Hand Action Recognition We compare our proposed method with existing state-of-the-art methods including Joule-color [17], Two Stream [9], H+O [44], Collaborative [52],

Table 2: 3D pose estimation performance in Root-Aligned space on the FPHA [15] and H2O [24]. We report AUC-RA for 3D PCK-RA at error thresholds ranging from 0 to 50 *mm* and the MEPE-RA in the unit of *mm*.

Model	AUC-RA(0-50) (↑)	MEPE-RA (↓)
HTT [49]	0.763	12.13
Ours	**0.769**	**11.79**

(a) FPHA Dataset

Model	AUC-RA(0-50) (↑)		MEPE-RA (↓)	
	Left	Right	Left	Right
HTT [49]	0.674	0.648	16.59	17.91
Ours	**0.686**	**0.662**	**15.96**	**17.08**

(b) H2O Dataset

Table 3: Ablative study of input features for Egocentric Action Module (EAM) on Hand Action Recognition Accuracy (%). We investigate the usage of the hand type feature in (a). Also, we analyze the effectiveness of each cue on the action recognition task in (b).

Hand Type	EAM Input	Text Embedding	Accuracy (↑)	
			FPHA [15]	H2O [24]
✓	-	-	93.74	85.95
✓	✓	-	94.26	87.60
✓	✓	✓	**95.13**	**89.67**

(a) Effect of Text-based Hand Type Feature

Input Feature for Egocentric Action Module				Accuracy (↑)	
Image Feature	Hand Pose	Object Label	Hand Type	FPHA [15]	H2O [24]
✓	✓	✓	-	94.09	86.36
✓	✓	-	✓	93.74	87.19
✓	-	✓	✓	94.61	86.78
✓	✓	✓	✓	**95.13**	**89.67**

(b) Effect of Hand Type Cue

HTT [49], and Trear [25] on the FPHA [15] dataset (see Table 1 (a)). On the H2O dataset, we compare ours with C2D [48], I3D [45], SlowFast [10], H+O [44], ST-GCN [51], TA-GCN [24] and HTT [49] (see Table 1 (b)). As reported in Table 1 (a) and (b), our method generally outperforms other methods with state-of-the-art accuracy (FPHA [15]: $\mathbf{95.13}\%$, H2O [24]: $\mathbf{89.67}\%$). These results emphasize the effectiveness of our method in understanding the interaction between hands and objects.

Hand Pose Estimation We also demonstrate considerable performance on 3D hand pose estimation, scoring AUC for 3D PCK at error thresholds ranging from 0 to 50 *mm* and the MEPE measured in *mm* within the Root-Aligned (RA) space. Our method estimates hand pose more precisely than the current state-of-the-art method [49] (see Table 2). These experimental results verify that the semantic knowledge of the hand type benefits not only the action recognition network but also the entire network, resulting in high precision in 3D hand pose estimation.

5.3 Ablation Study

Effect of Text-based Hand Type Feature In this section, we evaluate the variants of our method across hand type features. As shown in Table 3 (a), we investigate the usage of the hand type feature. The Egocentric Action Module (EAM) performs better when the predicted probability distribution feature vector from the hand type estimation network is used as input than when the hand type is simply predicted as the auxiliary network. Further, utilizing text embedding of hand type improves accuracy over using the predicted feature vector.

Effect of Hand Type Cue With four types of cues (image feature, hand pose, object, and hand type), we analyze the effect of each cue on action recognition with and without each cue as input of the egocentric action module. Specifically, on the FPHA [15] and the H2O [24] datasets, hand type cue effectively enhances accuracy, as shown in Table 3 (b). We finally observe that employing hand type as a semantic prior plays a key role in action recognition.

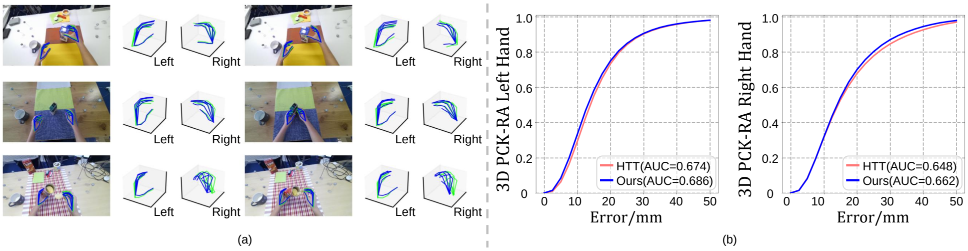

Figure 4: Qualitative result of our experiments. In (a), the green and blue line represents ground truth and estimated 3D hand pose, respectively. (b) shows the 3D PCK of hand pose estimation results on H2O [24] in Root-Aligned space. The blue line indicates the performance of our model, whereas the red line represents the HTT [49].

5.4 Qualitative Analysis

In this section, we qualitatively verify the usefulness of our novel framework. Fig. 4 shows the visualized results of our experiments. In Fig. 4 (a), the ground truth hand pose (see green lines) and our estimated hand pose (see blue lines) are projected in both 3D space and images. Ours show comparable results compared to ground truth. We also provide 3D PCK-RA graphs of left and right hand pose estimation results on the H2O test split of ours (see blue lines) vs. HTT [49] (see red lines) in Fig. 4 (b). This graph validates that our method generates reasonable results in both hands. Overall, our model is robust for estimating hand pose in 3D space. We provide more qualitative results and analysis in supplementary materials.

6 Conclusion

In this paper, we present a novel method applying the knowledge of hand type for hand action recognition based on the temporal transformer. This is the first attempt to regard the semantic details of the hand type as a critical indicator for enhancing the perception of egocentric view hand actions. To utilize the knowledge of hand type, we newly define the taxonomy based on hand functionality and annotate hand types for existing large-scale benchmarks. The experiments demonstrate the outstanding performance of our proposed approach.

Acknowledgement

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (23ZH1300, Research on hyper-realistic interaction technology for five senses and emotional experience, 95%), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00079, Artificial Intelligence Graduate School Program (Korea University), 3%), the National Research Foundation of Korea grant (NRF-2022R1F1A10 74334, 2%), and results of a study on the "Leaders in Industry-university Cooperation 3.0" Project, supported by the Ministry of Education and National Research Foundation of Korea. This work was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)& Gwangju Metropolitan City.

References

[1] Ian M Bullock, Júlia Borràs, and Aaron M Dollar. Assessing assumptions in kinematic hand models: a review. In *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics BioRob)*, pages 139–146. IEEE, 2012.

[2] Ian M Bullock, Thomas Feix, and Aaron M Dollar. Finding small, versatile sets of human grasps to span common objects. In *ICRA*. IEEE, 2013.

[3] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, volume 3. Ann Arbor, Michigan;, 2016.

[4] Minjie Cai, Kris M Kitani, and Yoichi Sato. An ego-vision system for hand grasp analysis. *IEEE Transactions on Human-Machine Systems*, 47(4):524–535, 2017.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[6] Mark R Cutkosky et al. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *IEEE Transactions on robotics and automation*, 5(3):269–279, 1989.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Alireza Fathi, Ali Farhadi, and James M Rehg. Understanding egocentric activities. In *ICCV*. IEEE, 2011.

[9] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016.

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

[11] Thomas Feix, Roland Pawlik, Heinz-Bodo Schmiedmayer, Javier Romero, and Danica Kragic. A comprehensive grasp taxonomy. In *Robotics, science and systems: workshop on understanding the human hand for advancing robotic manipulation*, volume 2, pages 2–3. Seattle, WA, USA;, 2009.

[12] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on human-machine systems*, 46(1):66–77, 2015.

[13] Qing Gao, Jinguo Liu, and Zhaojie Ju. Robust real-time hand detection and localization for space human–robot interaction based on deep learning. *Neurocomputing*, 390:198–206, 2020.

[14] Qing Gao, Yongquan Chen, Zhaojie Ju, and Yi Liang. Dynamic hand gesture recognition based on 3d hand pose estimation for human–robot interaction. *IEEE Sensors Journal*, 22(18):17421–17430, 2021.

[15] 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.

[16] René Gilster, Constanze Hesse, and Heiner Deubel. Contact points during multidigit grasping of geometric objects. *Experimental brain research*, 217:137–151, 2012.

[17] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *CVPR*, 2015.

[18] Thea Iberall. Human prehension and dexterous robot hands. *The International Journal of Robotics Research*, 16(3):285–299, 1997.

[19] Georgiana Juravle, Heiner Deubel, and Charles Spence. Attention and suppression affect tactile perception in reach-to-grasp movements. *Acta psychologica*, 138(2):302–310, 2011.

[20] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *CVPR*, 2017.

[21] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[22] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *CVPRW*. IEEE, 2017.

[23] Roberta L Klatzky, Brian McCloskey, Sally Doherty, James Pellegrino, and Terence Smith. Knowledge about hand shaping and knowledge about objects. *Journal of motor behavior*, 19(2):187–213, 1987.

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

[25] Xiangyu Li, Yonghong Hou, Pichao Wang, Zhimin Gao, Mingliang Xu, and Wanqing Li. Trear: Transformer-based rgb-d egocentric action recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):246–252, 2021.

[26] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *CVPR*, 2015.

[27] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV 2016*. Springer, 2016.

[28] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention lstm networks for 3d action recognition. In *CVPR*, 2017.

[29] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *ECCV*. Springer, 2020.

[30] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, 2018.

[31] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*. PMLR, 2022.

[32] T Manti. A novel type of compliant underactuated robotic hand for grasping. *Soft Robotics*, 35:161–185, 2015.

[33] Osama Mazhar, Benjamin Navarro, Sofiane Ramdani, Robin Passama, and Andrea Cherubini. A real-time human-robot interaction framework with robust background invariant hand gesture detection. *Robotics and Computer-Integrated Manufacturing*, 60:34–48, 2019.

[34] John R Napier. The prehensile movements of the human hand. *The Journal of bone and joint surgery. British volume*, 38(4):902–913, 1956.

[35] Supreeth Narasimhaswamy, Zhengwei Wei, Yang Wang, Justin Zhang, and Minh Hoai. Contextual attention for hand detection in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9567–9576, 2019.

[36] Yuzhe Qin, Hao Su, and Xiaolong Wang. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. *IEEE Robotics and Automation Letters*, 7(4):10873–10881, 2022.

[37] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *ECCV*. Springer, 2022.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.

[39] Georg Schlesinger. Der mechanische aufbau der künstlichen glieder. *Ersatzglieder und Arbeitshilfen: Für Kriegsbeschädigte und Unfallverletzte*, pages 321–661, 1919.

[40] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019.

[41] Roy Shilkrot13, Supreeth Narasimhaswamy, Saif Vazir, and Minh Hoai12. Workinghands: A hand-tool assembly dataset for image segmentation and activity mining. 2019.

[42] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *NeurIPS*, 27, 2014.

[43] Suriya Singh, Chetan Arora, and CV Jawahar. First person action recognition using deep learned descriptors. In *CVPR*, 2016.

[44] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, 2019.

[45] Quo Vadis, Joao Carreira, and Andrew Zisserman. Action recognition? a new model and the kinetics dataset. *Joao Carreira, Andrew Zisserman.*

[46] Limin Wang, Yu Qiao, Xiaoou Tang, et al. Action recognition and detection by combining motion and appearance features. *THUMOS14 Action Recognition Challenge*, 1 (2):2, 2014.

[47] Weitian Wang, Rui Li, Zachary Max Diekel, Yi Chen, Zhujun Zhang, and Yunyi Jia. Controlling object hand-over in human–robot collaboration via natural wearable sensing. *IEEE Transactions on Human-Machine Systems*, 49(1):59–71, 2018.

[48] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[49] 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.

[50] Cai-Hua Xiong, Wen-Rui Chen, Bai-Yang Sun, Ming-Jin Liu, Shi-Gang Yue, and Wen-Bin Chen. Design and implementation of an anthropomorphic hand for replicating human grasping functions. *IEEE Transactions on Robotics*, 32(3):652–671, 2016.

[51] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[52] Siyuan Yang, Jun Liu, Shijian Lu, Meng Hwa Er, and Alex C Kot. Collaborative learning of gesture recognition and 3d hand pose estimation with multi-order feature analysis. In *ECCV*. Springer, 2020.

[53] Yezhou Yang, Cornelia Fermuller, Yi Li, and Yiannis Aloimonos. Grasp type revisited: A modern perspective on a classical feature for vision. In *CVPR*, 2015.

[54] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, 2017.