First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations

Guillermo Garcia-Hernando Shanxin Yuan Seungryul Baek Tae-Kyun Kim
Imperial College London
{g.garcia-hernando,s.yuan14,s.baek15,tk.kim}@imperial.ac.uk

Abstract

In this work we study the use of 3D hand poses to recognize first-person dynamic hand actions interacting with 3D objects. Towards this goal, we collected RGB-D video sequences comprised of more than 100K frames of 45 daily hand action categories, involving 26 different objects in several hand configurations. To obtain hand pose annotations, we used our own mo-cap system that automatically infers the 3D location of each of the 21 joints of a hand model via 6 magnetic sensors and inverse kinematics. Additionally, we recorded the 6D object poses and provide 3D object models for a subset of hand-object interaction sequences. To the best of our knowledge, this is the first benchmark that enables the study of first-person hand actions with the use of 3D hand poses. We present an extensive experimental evaluation of RGB-D and pose-based action recognition by 18 baselines/state-of-the-art approaches. The impact of using appearance features, poses, and their combinations are measured, and the different training/testing protocols are evaluated. Finally, we assess how ready the 3D hand pose estimation field is when hands are severely occluded by objects in egocentric views and its influence on action recognition. From the results, we see clear benefits of using hand pose as a cue for action recognition compared to other data modalities. Our dataset and experiments can be of interest to communities of 3D hand pose estimation, 6D object pose, and robotics as well as action recognition.

1. Introduction

We interact with the world using our hands to manipulate objects, machines, tools, and socialize with other humans. In this work we are interested in understanding how we use our hands while performing daily life *dynamic* actions with the help of fine-grained hand pose features, a problem of interest for multiple applications requiring high precision, such as hand rehabilitation [1], virtual/augmented reality [24], teleoperation [89], and robot imitation learning [2].

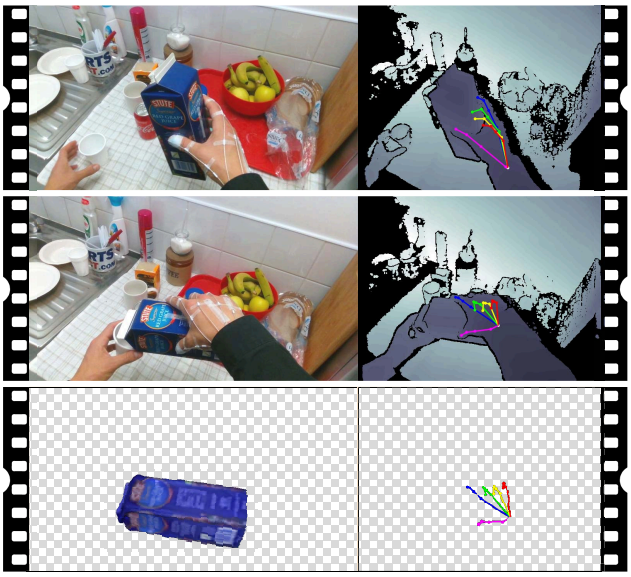

Figure 1: We show two frames from a sequence belonging to the action class 'pour juice'. We propose a novel first-person action recognition dataset with RGB-D videos and 3D hand pose annotations. We use magnetic sensors and inverse kinematics to capture the hand pose. On the right we see the captured depth image and hand pose. We also captured 6D object pose for a subset of hand-object actions.

Previous work in first-person action recognition [8, 23, 31, 56] found that daily actions are well explained by looking at hands, a similar observation found in third-person view [77]. In these approaches, hand information is extracted from hand silhouettes [31, 56] or discrete grasp classification [8, 23, 50] using low-level image features. In full-body human action recognition it is known that using higher level and viewpoint invariant features such as body pose can benefit action recognition [54, 74, 79, 81, 86], although this has not yet been studied in detail for hands. Compared to full-body actions, hand actions present unique differences that make the use of pose as a cue not obvious: style and

2575-7075/18 $31.00 © 2018 IEEE
DOI 10.1109/CVPR.2018.00050

IEEE computer society

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 03:18:54 UTC from IEEE Xplore. Restrictions apply.

speed variations across subjects are more pronounced due to a higher degree of mobility of fingers and the motion can be very subtle. A setback for using hand pose for action recognition is the absence of reliable pose estimators off-the-shelf in contrast to full body [55, 71], mainly due to the absence of hand pose annotations on real (*cf.* synthetic) data sequences, notably when objects are involved [10, 35, 48, 49].

In this work we introduce a new dataset of first-person dynamic hand action sequences with more than 100,000 RGB-D frames annotated with 3D hand poses, using six magnetic sensors attached to the fingertips and inverse kinematics. We captured 1175 action samples including 45 categories manipulating 26 different objects in 3 scenarios. We designed our hand actions and selected objects to cover multiple hand configurations and temporal dynamics. Furthermore, to encourage further research, we also provide 6-dimensional object pose ground truth, and their 3D mesh models, for 4 objects spanning, 10 different actions. We evaluate several baselines and state-of-the-art RGB-D and pose-based action recognition in our dataset and test the current state-of-the-art in hand pose estimation and its influence on action recognition. To the best of our knowledge, this is the first work that studies the problem of first-person action recognition with the use of hand pose features and the first benchmark of its kind. In summary, the contribution of this paper is three-fold:

Dataset: we propose a fully annotated dataset to help the study of egocentric dynamic hand-object actions and poses. This is the first dataset to combine both fields in the context of hands in real videos and quality hand pose labels.

Action recognition: we evaluate 18 baselines and state-of-the-art approaches in RGB-D and pose-based action recognition using our proposed dataset. Our selected methods cover most of the research trends in both methodology and use of different data modalities.

Hand pose: We evaluate a state-of-the-art hand pose estimator in our real dataset, *i.e.*, the occluded setting of hand-object manipulations and assess its performance for action recognition.

2. Related work

Egocentric vision and manipulations datasets: The important role of hands while manipulating objects has attracted the interest from both computer vision and robotics communities. From an action recognition perspective and only using RGB cues, recent research [5, 13, 14, 31, 44, 56] has delved into recognizing daily actions and determined that both manipulated objects and hands are important cues to the action recognition problem. A related line of work is the study of human grasp from a robotics perspective [6, 7], as a cue for action recognition [8, 16, 23, 77], force estimation [16, 26, 50], and as a recognition problem itself [20, 50]. Recently, [30] proposed a benchmark using a thermal camera enabling easier hand detection without exploring its use for action recognition. In these previous works, hands are modeled using low-level features or intermediate representations following empirical grasp taxonomies [6] and thus are limited compared to the 3D hand pose sequences used in this work. In [50], synthetic hand poses are used to recognize grasps in static frames, whereas our interest is in dynamic actions and hand poses in real videos. From a hand pose estimation perspective, [48] proposed a small synthetic dataset of static poses and thus could not succesfully train data-hungry algorithms, recently relieved by larger synthetic datasets [10, 35]. Given that we also provide 6D object poses and 3D mesh models for a subset of objects, our dataset can be of interest to both object pose and joint hand-object tracking emerging communities [57, 62]. We compare our dataset with other first-person view datasets in Section 3.5.

RGB-D and pose-based action recognition: Using depth sensors differs from traditional color action recognition in the fact that most successful color approaches [15, 67] cannot be directly applied to the depth stream due to its nature: noisy, textureless and discontinuous pixel regions led to the necessity of depth-tailored methods. These methods usually focus on how to extract discriminative features from the depth images using local geometric descriptors [40, 43, 76] sensitive to viewpoint changes and view-invariant approaches [46, 47]. However, the recent trend is to take advantage of the depth channel to obtain robust body pose estimates [55] and use them directly as a feature to recognize actions, what is known as pose or skeleton action recognition. Popular approaches include the use of temporal state-space models [17, 70, 74, 75, 86], key-poses [66, 85], hand-crafted pose features [64, 65], and temporal recurrent models [12, 63, 87]. Having multiple data streams has led to the study of combining different sources of information such as depth and pose [4, 40, 52, 69], color and pose [88], and all of them [19]. Most previous works in RGB-D action recognition focus on actions performed by the whole human body with some exceptions that are mainly application-oriented, such as hand gestures for human-computer interaction [9, 11, 29, 34, 40] and sign language [68]. Related to us, [33] mounted a depth sensor to recognize egocentric activities and modeling hands using low-level skin features. Similar to our interests but in third-person view, [27, 78] used a hand tracker to obtain noisy estimates of hand pose in kitchen manipulation actions, while [11] recognized basic hand gestures for human-computer interaction without objects involved. In these works, actions performed and pose labels are very limited due to the low quality of the hand tracker, while in this work we provide accurate hand pose labels to study more realistic hand actions. We go in depth and evaluate several baselines and state-of-the-art approaches in Sections 4 and 5.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 03:18:54 UTC from IEEE Xplore. Restrictions apply.

Figure 2: **Hand actions:** We captured daily hand actions using a RGB-D sensor and used a mo-cap system to annotate hand pose. **Left:** 'put sugar' and 'pour milk' (kitchen). **Right:** 'charge cell phone' (office) and 'handshake' (social).

3D hand pose estimation: Mainly due to the recent availability of RGB-D sensors, the field has made significant progress in object-less third-person view [22, 25, 28, 37, 39, 41, 45, 53, 60, 80] and more modest advances in first-person view [10, 35, 38, 48]. In [42], 3D tracking of a hand interacting with an object in third-person view was investigated. [18] studied the use of object-grasp as hand pose prior, while [51] used the object shape as cue. An important limitation is the difficulty of obtaining accurate 3D hand pose annotations leading researchers to resort to synthetic [3, 10, 35, 48, 53], manually or semi-automatically annotated [38, 58, 59, 61] datasets, resulting in non-realistic images, a low number of samples, and often inconsistent annotations. With the help of magnetic sensors for annotation and similar to [72], [84] proposed a big benchmark that included egocentric poses with no objects involved and showed that a ConvNet baseline can achieve state-of-the-art performance when enough training data is available. This was confirmed in a public challenge [83], also using a subset of our proposed dataset, and followed by a work [82] analyzing the current state-of-the-art of the field.

3. Daily hand-object actions dataset

3.1. Dataset overview

The dataset contains 1,175 action videos belonging to 45 different action categories, in 3 different scenarios, and performed by 6 actors. A total of 105,459 RGB-D frames are annotated with accurate hand pose and action category. Action sequences present high inter-subject and intra-subject variability of style, speed, scale, and viewpoint. The object's 6-dimensional pose, 3D location and angle, and mesh model are also provided for 4 objects involving 10 different action categories. Our plan is to keep growing the dataset with more models and objects. In Fig. 2 we show some example frames for different action categories and hand-pose annotation visualization.

3.2. Hand-object actions

We captured 45 different daily hand action categories involving 26 different objects. We designed our action categories to span a high number of different hand configurations following the same taxonomy as [50] and to be diverse in both hand pose and action space (see Fig. 4). Each object has a minimum of one associated action (*e.g.*, pen-'write') and a maximum of four (*e.g.*, sponge-'wash', 'scratch', 'squeeze', and flip'). These 45 hand actions were recorded and grouped in three different scenarios: kitchen (25), office (12) and social (8). In this work we consider each hand-object manipulation as a different action category similar to previous datasets [14], although other definitions are possible [73, 78].

3.3. Sensors and data acquisition

Visual data: We mounted an Intel RealSense SR300 RGB-D camera on the shoulder of the subject and captured sequences at 30 fps and resolutions 1920×1080 and 640×480 for the color and depth stream respectively.

Pose annotation: To obtain quality annotations of hand and object pose, the hand pose is captured using six magnetic sensors [36] attached to the user's hand, five fingertips and one wrist, following [84]. Each sensor provides position and orientation with 6 degrees of freedom and the full hand pose is inferred using inverse kinematics over a defined 21-joint hand model. Each sensor is 2 mm wide and when attached to the human hand does not influence the depth image. The color image is affected as the sensors and the tape attaching them are visible, however the hand is fully visible and actions distinguishable by using the color image. Regarding object pose, we attach one more sensor to the closest point to the center of mass that can be reached.

Recording process: We asked 6 people, all right-handed, to perform the actions. Instructions on how to perform the action in a safe manner were given, however no instructions about style or speed were provided, in order to capture realistic data. Actions were labeled manually.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 03:18:54 UTC from IEEE Xplore. Restrictions apply.

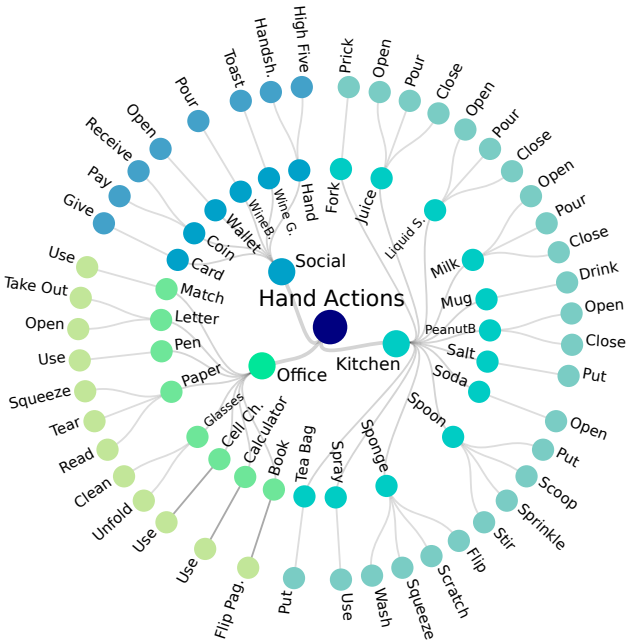

Figure 3: Taxonomy of our hand actions involving objects dataset. Some objects are associated with multiple actions (*e.g.*, spoon, sponge, liquid soap), while some others have only one linked action (*e.g.*, calculator, pen, cell charger).

3.4. Dataset statistics

Taxonomy: Fig. 3 shows the distribution of different actions per involved object. Some objects such as 'spoon' have multiple actions (*e.g.*, 'stir', 'sprinkle', 'scoop', and 'put sugar'), while some objects have only one action ('use calculator'). Although it is not an object per se, we included 'hand' as an object in actions 'handshake' and 'high five'.

Videos per action class: On average there are 26.11 sequences per class action and 45.19 sequences per object. For detailed per class numbers see Fig. 4 (c).

Duration of videos: Fig. 4 (d) shows the average number of video duration for the 45 action classes. Some action classes such as 'put sugar' and 'open wallet' involve short atomic movements, on average one second, while others such as 'open letter' require more time to be executed.

Grasps: We identified 34 different grasps following the same taxonomy as in [50], including the most frequently studied ones [8] (*i.e.*, precision/power grasps for different object attributes such as prismatic/round/flat/deformable). In Fig. 4 (b) we show some examples of correlation between objects, hand poses, and actions.

Viewpoints: In Fig. 4 (e) we show the distribution of frames per hand viewpoint. We define the viewpoint as the angle between the camera direction and the palm of the hand. The dataset presents viewpoints that are more prone to self-occlusion than typical ones in third-person view.

Dataset	Sensor	Real?	Class.	Seq.	Frames	Labels
Yale [6]	RGB	✓	33	-	9,100	Grasp
UTG [7]	RGB	✓	17	-	-	Grasp
GTEA [14]	RGB	✓	61	525	31,222	Action
EgoHands [5]	RGB	✓	4	48	4,800	Action
GUN-71 [50]	RGB-D	✓	71	-	12,000	Grasp
UCI-EGO [48]	RGB-D	✗	-	-	400	Pose
Choi *et al.* [10]	RGB-D	✗	33	-	16,500	Grasp+Pose
SynthHands [35]	RGB-D	✗	-	-	63,530	Pose
EgoDexter [35]	RGB-D	✓	-	-	3,190	Fingertips
Luo *et al.* [30]	RGB-D-T	✓	44	250	450,000	Action
Ours	RGB-D	✓	45	1,175	105,459	Action+Pose

Table 1: First-person view datasets with hands and objects involved. Our proposed dataset is the first providing both hand pose and action annotations in real data (*cf.* synthetic).

Hand occlusion: Fig. 6 (a) (bottom) shows the average number of visible (not occluded by object or viewpoint) hand joints per action class. Most actions present a high degree of occlusion, on average 10 visible joints out of 21.

Object pose: 6D object pose and mesh models are provided for the following objects involving 10 different actions: 'milk bottle', 'salt', 'juice carton', and 'liquid soap'.

3.5. Comparison with other datasets

In Table 1 we summarize popular egocentric datasets that involve hands and objects in both dynamic and static fashion depending on their problem of interest. For conciseness, we have excluded from the table related datasets that do not partially or fully contain objects manipulations, *e.g.*, [38, 44, 84]. Note that previous datasets in action recognition [5, 14, 30] do not include hand pose labels. On the other hand, pose and grasp datasets [6, 7, 10, 35, 48, 50] do not contain dynamic actions and hand pose annotation is obtained by generating synthetic images or rough manual annotations [35]. Our dataset 'fills the gap' of egocentric dynamic hand action using pose and compares favorably in terms of diversity, number of frames, and use of real data.

4. Evaluated algorithms and baselines

4.1. Action recognition

In order to evaluate the current state-of-the-art in action recognition we chose a variety of approaches that, we believe, cover the most representative trends in the literature as shown in Table 4. As the nature of our data is RGB-D and we have hand pose, we focus our attention to RGB-D and pose-based action recognition approaches, although we also evaluate two RGB action recognition methods [15, 19]. Note that, as discussed above, most of previous works in RGB-D action recognition involve full body poses instead of hands and some of them might not be tailored for hand actions. We elaborate further on this in Section 5.1.

We start with one baseline to assess how the current state-of-the-art in RGB action recognition performs in our dataset. For this, and given that most successful RGB action

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 03:18:54 UTC from IEEE Xplore. Restrictions apply.

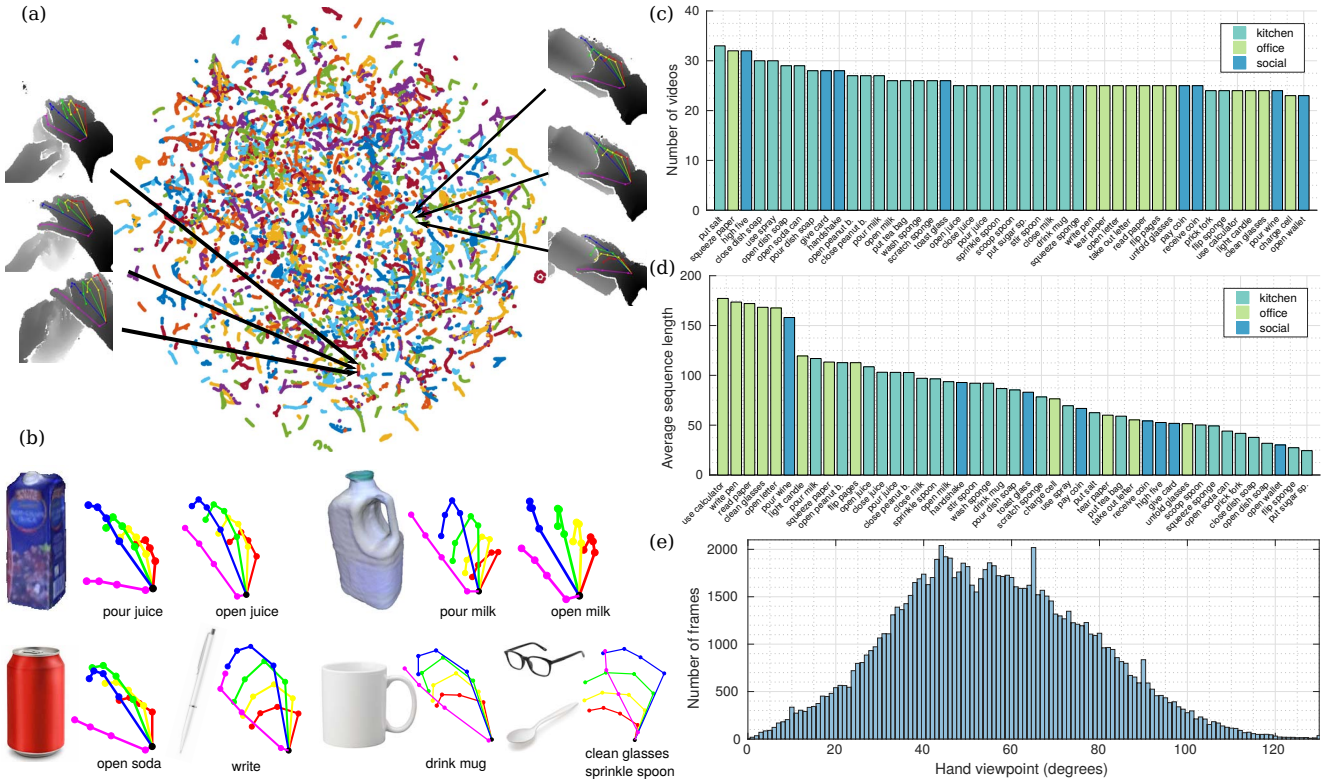

Figure 4: **(a)** t-SNE [32] visualization of hand pose embedding over our dataset. Each colored dot represents a full hand pose and each trajectory an action sequence. **(b)** Correlation between objects, grasps, and actions. Shown poses are the average pose over all action sequences of a certain class. One object can have multiple grasps associated depending on the action performed (*e.g.*, 'juice carton' and 'milk bottle') and one grasp can have multiple actions associated (*e.g.*, lateral grasp present at 'sprinkle' and 'clean glasses'). **(c)** Number of action instances per hand action class. **(c)** Average number of frames in each video per hand action class. Our dataset contains both atomic and more temporally complex action classes. **(d)** Distribution of hand viewpoints, defined as angles between the direction of the camera and the direction of the palm of the hand.

recognition approaches [31, 56] use ConvNets to learn descriptors from color and motion flow, we evaluate a recent two-stream architecture fine-tuned on our dataset [15].

About the depth modality, we first evaluate two local depth descriptor approaches, HOG^2 [40] and HON4D [43], that exploit gradient and surface normal information as a feature for action recognition. As a global-scene depth descriptor, we evaluate the recent approach by [47] that learns view invariant features using ConvNets from several synthesized depth views of human body pose.

We follow our evaluation with pose-based action recognition methods. As our main baseline, we implemented a recurrent neural network with long-short term memory (LSTM) modules inspired in the architecture by [87]. We also evaluate several state-of-the-art pose action recognition approaches. We start with descriptor-based methods such as Moving Pose [85] that encodes atomic motion information and [64], which represents poses as points on a Lie group. For methods focusing on learning temporal dependencies, we evaluate HBRNN [12], Gram Matrix [86] and TF [17]. HBRNN consists of a bidirectional recurrent neural network with hierarchical layers designed to learn features from the body pose. Gram Matrix is currently the best performing method for body pose and uses Gram matrices to learn the dynamics of actions. TF learns both discriminative static poses and transitions between poses using decision forests.

To conclude, we evaluate one hybrid approach that jointly learns heterogeneous features (JOULE) [19] using an iterative algorithm to learn features jointly taking into account all the data channels: color, depth, and hand pose.

4.2. Hand pose estimation

To assess the state-of-the-art in hand pose estimation, we use the same ConvNet as [84]. We choose this approach as it is easy to interpret and it was shown to provide good performance in a cross-benchmark evaluation [84]. The chosen method is a discriminative approach operating on a frame-by-frame basis, which does not need any initialization and manual recovery when tracking fails [21, 41].

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 03:18:54 UTC from IEEE Xplore. Restrictions apply.

5. Benchmark evaluation results

5.1. Action recognition

In the following we present our experiments in action recognition. In this section we assume the hand pose is given, *i.e.*, we use the hand pose annotations obtained using the magnetic sensors and inverse-kinematics. We evaluate the use of estimated hand poses without the aid of the sensors for action recognition in Section 5.2.

Following common practice in full body-pose action recognition [64, 85], we compensate for anthropomorphic and viewpoint differences by normalizing poses to have the same distance between pairs of joints and defining the wrist as the center of coordinates.

5.1.1 A baseline: LSTM

We start our experimental evaluation with a simple yet powerful baseline: a recurrent neural network with long-short term memory module (LSTM). The architecture of our network is inspired by [87] with two differences: we do not 'go deep', and use a more conventional unidirectional network instead of bidirectional. Following [87], we set the number of neurons to 100 and a probability of dropout of 0.2. We use TensorFlow and Adam optimizer.

Training and testing protocols: We experiment with two protocols. The first protocol consists of using different partitions of the data for training and the rest for testing and we tried three different training:testing ratios of 1:3, 1:1 and 3:1 at sequence level. The second protocol is a 6-fold 'leave-one-person-out' cross-validation, *i.e.*, each fold consists of 5 subjects for training and one for testing. Results are presented in Table 2. We observe that following a cross-person protocol yields the worst results taking into account that in each fold we have similar training/testing proportions to the $3:1$ setting. This can be explained by the difference in hand action styles between subjects. In the rest of the paper we perform our experiments using the 1:1 setting with 600 action sequences for training and 575 for testing.

Results discussion: In Fig. 5 (a) we show the recognition accuracies per category on a subset actions and the action confusion matrix is shown in Fig. 6 (b). Some actions such as 'sprinkle spoon', 'put tea bag' and 'pour juice' are easily identifiable, while actions such as 'open wallet' and 'use calculator' are commonly confused, likely because hand poses are dissimilar and more subtle. In Fig. 5 (d) we show the contribution of each finger motion to action recognition performance, finding that the index is the most informative finger. Combining thumb and index poses boosts the accuracy, likely due to the fact that most grasps are explained by these two fingers [6]. Fingertips alone are also a high source of information due to being the highest articulated joints and being able to 'explain' the hand pose.

Protocol	1:3	1:1	3:1	cross-person
Acc. (%)	58.75	78.73	84.82	62.06

Table 2: Action recognition results (percentage of correct video classification) for different training/testing protocols.

5.1.2 State-of-the-art evaluation

In Table 4 we show results for state-of-the-art approaches in different data modalities. We observe that the Two-stream [15] method performs well when combining both spatial and temporal cues. Depth methods tend to perform slightly worse than the rest of the methods, suggesting that they are not able to fully capture either the object cues or the hand pose. Note that for Novel View [47] we extracted deep features from a network trained on several synthetic views of bodies, which may not generalize well to hand poses and fine-tuning in our dataset did not help. From all approaches, we observe that the ones using hand pose are the ones that achieve the best performance, with Gram Matrix [86] and Lie group [64] performing particularly well, a result in line with the ones reported in body pose action recognition.

In Fig. 5 we select some of the most representative methods and analyze their performance in detail. We observe that the pose method Gram Matrix outperforms the rest in most of the measures, specially when we retrieve the top k action hypothesis (Fig. 5 (b)), showing the benefit of using hand pose for action recognition. Looking at Fig. 5 (a), we observe that Two-stream outperforms the rest of methods in some categories in which the object is big and the action does not involve much motion, *e.g.*, 'use calculator' and 'read paper'. This good performance can be due to the pre-training of the spatial network on a big image recognition dataset. We further observe this in Fig. 5 (c) where we analyze the top k hypothesis given by the prediction and look whether the predicted action contains the object being manipulated, suggesting that the network correctly recognizes the object but fails to capture the temporal dynamics.

Hand pose vs. depth vs. color: We performed one additional experiment using the JOULE [19] approach by breaking down the contributions of each data modality. In Table 4 (bottom) we show that hand pose features are the most discriminative ones, although the performance can be increased by combining them with RGB and depth cues. This result suggests that hand poses capture complementary information to RGB and depth features.

Object pose: We did an additional experiment using the object pose as a feature for action recognition using the subset of actions that have annotated object poses: a total of 261 sequences for 10 different classes and 4 objects. We trained our LSTM baseline on half of the sequences and using three different inputs: hand pose, object pose, and both combined. In Table 3 we show the results and observe that

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 03:18:54 UTC from IEEE Xplore. Restrictions apply.

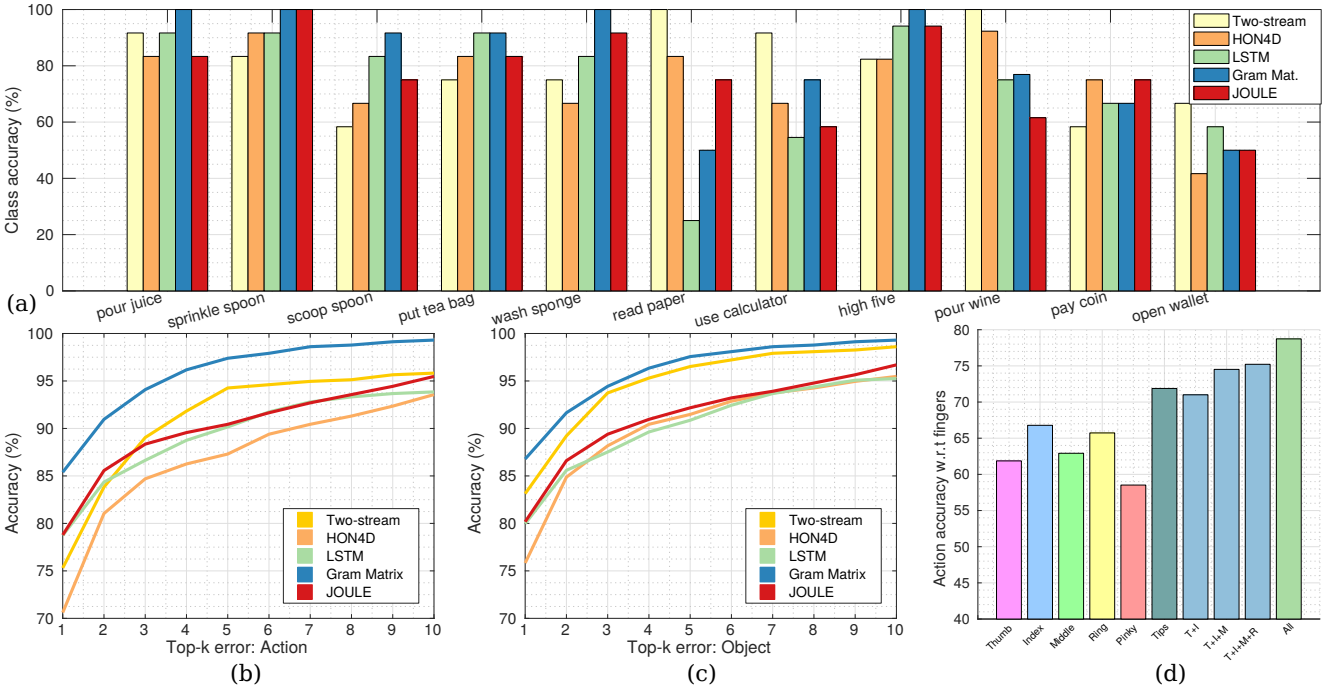

Figure 5: **(a)** Class accuracies of some representative methods on a subset of classes. **(b)** Top-k action accuracy: true action label is in the top-k action prediction hypothesis. **(c)** Top-k object accuracy: manipulated object is in the top-k action prediction hypothesis. **(d)** Impact of each of the five fingers, combinations of them, and fingertips on action recognition.

Pose feature	Hand	Object	Hand+Object
Action acc. (%)	87.45	74.45	91.97

Table 3: We evaluate the use of 6D object pose for action recognition on a subset of our dataset. We observe the benefit of combining them with the hand pose.

both object pose and hand pose features are complimentary and useful for recognizing egocentric hand-object actions.

5.2. Hand pose estimation

Training with objects vs. no objects: One question raised while designing our experiments was whether we actually needed to annotate the hand pose in a close to ground truth accuracy to experiment with hand dynamic actions. We try to answer this question by estimating the hand poses of our hand action dataset in two ways partitioning our data as in our *Action split*: using the nearly 300k *object-free* egocentric samples from [84] and using the images in the training set of our hand action dataset. As observed in Fig. 6 (c) and Table 5, the results suggest that having hand-object images in the training set is crucial to train state-of-the-art hand pose estimators, likely due to the fact that occlusions and object shapes need to be *seen* by the estimator beforehand. To confirm this, we conducted two extra experiments: *cross-subject* (half of the users in training and half in testing, all objects seen in both splits) and *cross-object* (half of the objects in training and half in testing, all subjects seen in both splits). In Fig. 6 (c) and Table 5 we observe that the network is able to generalize to unseen subjects but struggles to do so for unseen objects, suggesting that recognizing the shape of the object and its associated grasp is crucial to train hand pose estimators. This shows the need of having

Method	Year	Color	Depth	Pose	Acc. (%)
Two stream-color [15]	2016	✓	✗	✗	61.56
Two stream-flow [15]	2016	✓	✗	✗	69.91
Two stream-all [15]	2016	✓	✗	✗	75.30
HOG^2-depth [40]	2013	✗	✓	✗	59.83
HOG^2-depth+pose [40]	2013	✗	✓	✓	66.78
HON4D [43]	2013	✗	✓	✗	70.61
Novel View [47]	2016	✗	✓	✗	69.21
1-layer LSTM	2016	✗	✗	✓	78.73
2-layer LSTM	2016	✗	✗	✓	80.14
Moving Pose [85]	2013	✗	✗	✓	56.34
Lie Group [64]	2014	✗	✗	✓	82.69
HBRNN [12]	2015	✗	✗	✓	77.40
Gram Matrix [86]	2016	✗	✗	✓	85.39
TF [17]	2017	✗	✗	✓	80.69
JOULE-color [19]	2015	✓	✗	✗	66.78
JOULE-depth [19]	2015	✗	✓	✗	60.17
JOULE-pose [19]	2015	✗	✗	✓	74.60
JOULE-all [19]	2015	✓	✓	✓	78.78

Table 4: Hand action recognition performance by different evaluated approaches on our proposed dataset.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 03:18:54 UTC from IEEE Xplore. Restrictions apply.

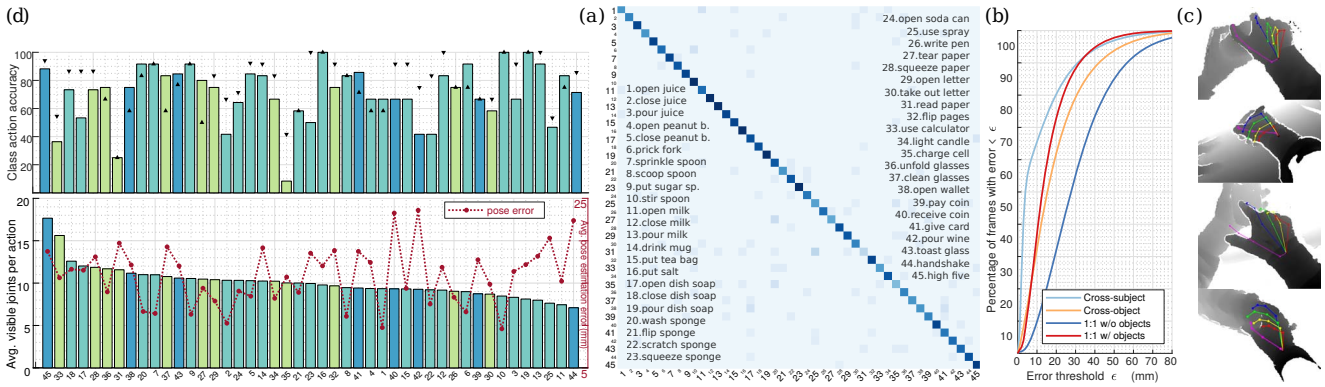

Figure 6: **(a)** *Top:* Class action recognition accuracies for our LSTM baseline using estimated hand poses (accuracies with groundtruth poses are represented with black triangles). *Bottom*: Average number of visible (not occluded) joints for hand actions on our dataset and its impact on hand pose estimation. **(b)** Hand action confusion matrix for our LSTM baseline. **(c)** Percentage of frames for different hand pose estimation error thresholds. **(d)** Qualitative results on hand pose estimation.

annotated hand poses interacting with objects and thus why our dataset can be of interest for the hand pose community. In Fig. 6 (d) we show some qualitative results in hand pose estimation in our proposed dataset and observe that, while not perfect, they are good enough for action recognition.

Hand pose estimation and action recognition: Now we try to answer the following key question: 'how good is the current hand pose estimation for recognizing hand actions?'. In Table 5 we show results of hand action recognition by swapping the hand pose labels by the estimated ones in the test set. We observe that reducing the hand pose error by a factor of two yields a more than twofold improvement in action recognition. The difference in hand action recognition between using the hand pose labels and using the estimated ones in testing is 6.67%. We also tested the two best performant methods from previous section, Lie group [64] and Gram Matrix [86]. For Lie group we obtained an accuracy of 69.22%, while for Gram Matrix a poor result of 32.22% likely due to their strong assumptions in the noise distribution. On the other hand, our LSTM baseline shows more robust behavior in the presence of noisy hand pose estimates. In Fig. 6 (a) we show how the hand occlusion affects the pose estimation quality and its impact on class recognition accuracies. Although some classes present a clear correlation between pose error and action accuracy degradation (*e.g.*, 'receive coin', 'pour wine'), the LSTM is still able to obtain acceptable recognition rates likely due to being able to infer the action from temporal patterns and correctly estimated joints. For more insight, we analyzed the pose error per finger: T: 12.45, I: 15.48, M: 18.08, R: 16.69, P: 18.95, all in mm. Thumb and index joints present the lowest estimation error because of typically being less occluded in egocentric setting. According to previous section where we found that the motion from these two fingers was a high source of information, this can be a plausible explanation of why we can still obtain a good action recognition performance while having noisy hand pose estimates.

Hand pose protocol	Pose error (mm)	Action (%)
Cross-subject	11.25	-
Cross-object	19.84	-
Action split (training w/o objects)	31.03	29.63
Action split (training w/ objects)	14.34	72.06
Action split (GT mag.+IK poses)	-	78.73

Table 5: Average hand pose estimation error, 3D distance over all 21 joints between magnetic poses and estimates, for different protocols and its impact on action recognition.

6. Concluding remarks

We have proposed a novel benchmark and presented experimental evaluations for RGB-D and pose-based, hand action recognition, in first-person setting. The benchmark provides both temporal action labels and full 3D hand pose labels, and additionally 6D object pose labels on a part of the dataset. Both RGB-D action recognition and 3D hand pose estimation are relatively new fields, and this is a first attempt to relate both of them similar to full human body. We have evaluated several baselines in our dataset and concluded that hand pose features are a rich source of information for recognizing manipulation actions. We believe that our dataset and experiments can encourage future work in multiple fields including action recognition, hand pose estimation, object pose estimation, and emerging ones, such as joint hand-object pose estimation.

Acknowledgements: This work is part of Imperial College London-Samsung Research project, supported by Samsung Electronics.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 03:18:54 UTC from IEEE Xplore. Restrictions apply.

References

[1] S. Allin and D. Ramanan. Assessment of post-stroke functioning using machine vision. In *MVA*, 2007. 1

[2] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *RAS*, 2009. 1

[3] S. Baek, K. I. Kim, and T.-K. Kim. Augmented skeleton space transfer for depth-based hand pose estimation. In *CVPR*, 2018. 3

[4] S. Baek, Z. Shi, M. Kawade, and T.-K. Kim. Kinematic-layout-aware random forests for depth-based action recognition. In *BMVC*, 2017. 2

[5] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*, 2015. 2, 4

[6] I. M. Bullock, T. Feix, and A. M. Dollar. The yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. *IJRR*, 2015. 2, 4, 6

[7] M. Cai, K. M. Kitani, and Y. Sato. A scalable approach for understanding the visual structures of hand grasps. In *ICRA*, 2015. 2, 4

[8] M. Cai, K. M. Kitani, and Y. Sato. Understanding hand-object manipulation with grasp types and object attributes. In *RSS*, 2016. 1, 2, 4

[9] H. J. Chang, G. Garcia-Hernando, D. Tang, and T.-K. Kim. Spatio-temporal hough forest for efficient detection–localisation–recognition of fingerwriting in egocentric camera. *CVIU*, 2016. 2

[10] C. Choi, S. H. Yoon, C.-N. Chen, and K. Ramani. Robust hand pose estimation during the interaction with an unknown object. In *ICCV*, 2017. 2, 3, 4

[11] Q. De Smedt, H. Wannous, and J.-P. Vandeborre. Skeleton-based dynamic hand gesture recognition. In *CVPRW*, 2016. 2

[12] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 2015. 2, 5, 7

[13] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011. 2

[14] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011. 2, 3, 4

[15] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 2, 4, 5, 6, 7

[16] C. Fermüller, F. Wang, Y. Yang, K. Zampogiannis, Y. Zhang, F. Barranco, and M. Pfeiffer. Prediction of manipulation actions. *IJCV*, 2017. 2

[17] G. Garcia-Hernando and T.-K. Kim. Transition forests: Learning discriminative temporal transitions for action recognition and detection. In *CVPR*, 2017. 2, 5, 7

[18] H. Hamer, J. Gall, T. Weise, and L. Van Gool. An object-dependent hand pose prior from sparse training data. In *CVPR*, 2010. 3

[19] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for RGB-D activity recognition. In *CVPR*, 2015. 2, 4, 5, 6, 7

[20] D.-A. Huang, M. Ma, W.-C. Ma, and K. M. Kitani. How do we use our hands? discovering a diverse set of common grasps. In *CVPR*, 2015. 2

[21] Intel. Perceptual computing sdk. 2013. 5

[22] C. Ionescu, J. Carreira, and C. Sminchisescu. Iterated second-order label sensitive pooling for 3d human pose estimation. In *CVPR*, 2014. 3

[23] T. Ishihara, K. M. Kitani, W.-C. Ma, H. Takagi, and C. Asakawa. Recognizing hand-object interactions in wearable camera videos. In *ICIP*, 2015. 1, 2

[24] Y. Jang, S.-T. Noh, H. J. Chang, T.-K. Kim, and W. Woo. 3d finger cape: Clicking action and position estimation under self-occlusions in egocentric viewpoint. *TVCG*, 2015. 1

[25] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *ECCV*, 2012. 3

[26] P. G. Kry and D. K. Pai. Interaction capture and synthesis. *ACM Transactions on Graphics*, 2006. 2

[27] J. Lei, X. Ren, and D. Fox. Fine-grained kitchen activity recognition using rgb-d. In *Ubicomp*, 2012. 2

[28] H. Liang, J. Yuan, and D. Thalmann. Parsing the hand in depth images. *TMM*, 2014. 3

[29] L. Liu and L. Shao. Learning discriminative representations from rgb-d video data. In *IJCAI*, 2013. 2

[30] R. Luo, O. Sener, and S. Savarese. Scene semantic reconstruction from egocentric rgb-d-thermal videos. In *3D Vision (3DV)*, 2017. 2, 4

[31] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *CVPR*, 2016. 1, 2, 5

[32] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 2008. 5

[33] M. Moghimi, P. Azagra, L. Montesano, A. C. Murillo, and S. Belongie. Experiments on an rgb-d wearable vision system for egocentric activity recognition. In *CVPRW*, 2014. 2

[34] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In *CVPR*, 2016. 2

[35] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *ICCV*, 2017. 2, 3, 4

[36] NDItrakSTAR. http://www.ascension-tech.com/products/trakstar-2-drivebay-2/. 3

[37] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. Hand segmentation with structured convolutional learning. In *ACCV*, 2014. 3

[38] M. Oberweger, G. Riegler, P. Wohlhart, and V. Lepetit. Efficiently creating 3d training data for fine hand pose estimation. In *CVPR*, 2016. 3, 4

[39] M. Oberweger, P. Wohlhart, and V. Lepetit. Training a feedback loop for hand pose estimation. In *ICCV*, 2015. 3

[40] E. Ohn-Bar and M. M. Trivedi. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *ITS*, 2014. 2, 5, 7

[41] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, 2011. 3, 5

[42] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *ICCV*, 2011. 3

[43] O. Oreifej and Z. Liu. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences. In *CVPR*, 2013. 2, 5, 7

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 03:18:54 UTC from IEEE Xplore. Restrictions apply.

[44] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 2, 4

[45] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. Realtime and robust hand tracking from depth. In *CVPR*, 2014. 3

[46] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian. Histogram of oriented principal components for cross-view action recognition. *TPAMI*, 2016. 2

[47] H. Rahmani and A. Mian. 3d action recognition from novel viewpoints. In *CVPR*, 2016. 2, 5, 6, 7

[48] G. Rogez, M. Khademi, J. Supančič III, J. M. M. Montiel, and D. Ramanan. 3d hand pose detection in egocentric rgb-d images. In *ECCV*, 2014. 2, 3, 4

[49] G. Rogez, J. S. Supancic, and D. Ramanan. First-person pose recognition using egocentric workspaces. In *CVPR*, 2015. 2

[50] G. Rogez, J. S. Supancic, and D. Ramanan. Understanding everyday hands in action from rgb-d images. In *ICCV*, 2015. 1, 2, 3, 4

[51] J. Romero, H. Kjellström, C. H. Ek, and D. Kragic. Non-parametric hand pose estimation with object context. *IVC*, 2013. 3

[52] A. Shahroudy, T.-T. Ng, Q. Yang, and G. Wang. Multimodal multipart learning for action recognition in depth videos. *TPAMI*, 2016. 2

[53] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. Fitzgibbon, and S. Izadi. Accurate, robust, and flexible real-time hand tracking. In *CHI*, 2015. 3

[54] Z. Shi and T.-K. Kim. Learning and refining of privileged information-based rnns for action recognition from depth sequences. In *CVPR*, 2017. 1

[55] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Commun. ACM*, 2013. 2

[56] S. Singh, C. Arora, and C. Jawahar. First person action recognition using deep learned descriptors. In *CVPR*, 2016. 1, 2, 5

[57] S. Sridhar, F. Mueller, M. Zollhöfer, D. Casas, A. Oulasvirta, and C. Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *ECCV*, 2016. 2

[58] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *CVPR*, 2015. 3

[59] D. Tang, H. J. Chang, A. Tejani, and T.-K. Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *CVPR*, 2014. 3

[60] D. Tang, T.-H. Yu, and T.-K. Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *ICCV*, 2013. 3

[61] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *TOG*, 2014. 3

[62] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 2016. 2

[63] V. Veeriah, N. Zhuang, and G.-J. Qi. Differential recurrent neural networks for action recognition. In *ICCV*, 2015. 2

[64] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, 2014. 2, 5, 6, 7, 8

[65] R. Vemulapalli and R. Chellappa. Rolling rotations for recognizing human actions from 3d skeletal data. In *CVPR*, 2016. 2

[66] C. Wang, Y. Wang, and A. L. Yuille. Mining 3d key-pose-motifs for action recognition. In *CVPR*, 2016. 2

[67] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 2

[68] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In *ECCV*, 2012. 2

[69] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012. 2

[70] P. Wang, C. Yuan, W. Hu, B. Li, and Y. Zhang. Graph based skeleton motion representation and similarity measurement for action recognition. In *ECCV*, 2016. 2

[71] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2

[72] A. Wetzler, R. Slossberg, and R. Kimmel. Rule of thumb: Deep derotation for improved fingertip detection. In *BMVC*, 2015. 3

[73] M. Wray, D. Moltisanti, W. Mayol-Cuevas, and D. Damen. Sembed: Semantic embedding of egocentric action videos. In *ECCW*, 2016. 3

[74] D. Wu and L. Shao. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *CVPR*, 2014. 1, 2

[75] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *CVPRW*, 2012. 2

[76] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *CVPR*, 2014. 2

[77] Y. Yang, C. Fermuller, Y. Li, and Y. Aloimonos. Grasp type revisited: A modern perspective on a classical feature for vision. In *CVPR*, 2015. 1, 2

[78] Y. Yang, A. Guha, C. Fermuller, and Y. Aloimonos. A cognitive system for understanding human manipulation actions. *ACS*, 2014. 2, 3

[79] A. Yao, J. Gall, G. Fanelli, and L. J. Van Gool. Does human action recognition benefit from pose estimation?. In *BMVC*, 2011. 1

[80] Q. Ye, S. Yuan, and T.-K. Kim. Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In *ECCV*, 2016. 3

[81] T.-H. Yu, T.-K. Kim, and R. Cipolla. Unconstrained monocular 3d human pose estimation by action detection and cross-modality regression forest. In *CVPR*, 2013. 1

[82] S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Y. Chang, K. M. Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge, J. Yuan, X. Chen, G. Wang, F. Yang, K. Akiyama, Y. Wu, Q. Wan, M. Madadi, S. Escalera, S. Li, D. Lee, I. Oikonomidis, A. Argyros, and T.-K. Kim. Depth-based 3d hand pose estimation: From current achievements to future goals. In *CVPR*, 2018. 3

[83] S. Yuan, Q. Ye, G. Garcia-Hernando, and T.-K. Kim. The 2017 hands in the million challenge on 3d hand pose estimation. *arXiv preprint arXiv:1707.02237*, 2017. 3

[84] S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim. Big hand 2.2m benchmark: Hand pose data set and state of the art analysis. In *CVPR*, 2017. 3, 4, 5, 7

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 03:18:54 UTC from IEEE Xplore. Restrictions apply.

[85] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *ICCV*, 2013. 2, 5, 6, 7

[86] X. Zhang, Y. Wang, M. Gou, M. Sznaier, and O. Camps. Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold. In *CVPR*, 2016. 1, 2, 5, 6, 7, 8

[87] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *AAAI*, 2016. 2, 5, 6

[88] Y. Zhu, W. Chen, and G. Guo. Fusing spatiotemporal features and joints for 3d action recognition. In *CVPRW*, 2013. 2

[89] L. Zollo, S. Roccella, E. Guglielmelli, M. C. Carrozza, and P. Dario. Biomechatronic design and control of an anthropomorphic artificial hand for prosthetic and robotic applications. *TMECH*, 2007. 1

Authorized licensed use limited to: ShanghaiTech University. Downloaded on September 11,2026 at 03:18:54 UTC from IEEE Xplore. Restrictions apply.