EPIC-Tent: An Egocentric Video Dataset for Camping Tent Assembly

Youngkyoon Jang Brian Sullivan Casimir Ludwig Iain D. Gilchrist Dima Damen
Walterio Mayol-Cuevas
University of Bristol
Bristol, UK

Abstract

This paper presents an outdoor video dataset annotated with action labels, collected from 24 participants wearing two head-mounted cameras (GoPro and SMI eye tracker) while assembling a camping tent. In total, this is 5.4 hours of recordings. Tent assembly includes manual interactions with non-rigid objects such as spreading the tent, securing guylines, reading instructions, and opening a tent bag. An interesting aspect of the dataset is that it reflects participants' proficiency in completing or understanding the task. This leads to participant differences in action sequences and action durations. Our dataset, called EPIC-Tent[1], also has several new types of annotations for two synchronised egocentric videos. These include task errors, self-rated uncertainty and gaze position, in addition to the task action labels. We present baseline results on the EPIC-Tent dataset using a state-of-the-art method for offline and online action recognition and detection.

Figure 1. Participant standing next to a completed tent, wearing the GoPro and SMI eye tracker. To avoid the sun's infrared interference, the eye tracker was outfitted with protective lenses and black cloth. (Top Left) Egocentric perspective from the SMI camera, the orange circle represents the participant's point of gaze. (Bottom Left) Egocentric perspective from the GoPro camera.

1. Introduction

We present a novel egocentric dataset, EPIC-Tent, intended for activity recognition on a challenging, outdoor and non-rigid objects scenario – camping tent assembly. Egocentric activity recognition offers a range of challenges as the camera moves in the world and observes complex object interactions. But it also allows for observations that are near the action and the purpose of the acting agent.

Various tasks have recently been considered in egocentric computer vision, such as handled object detection [10, 7], food recognition [8], socialising pattern characterisation [41] and event sequence description [6] in various environments including a kitchen [9], office [1] and outdoors [44]. Egocentric datasets are mostly focused on rigid or low Degrees of Freedom (DoF) articulated objects and or well-defined interactions. We are here interested in pushing the boundaries of egocentric video activity recognition by the introduction of a novel egocentric dataset consisting of participants assembling a camping tent. Annotated with the action and task error labels, participants' self-rated level of step uncertainty, and eye gaze position, EPIC-Tent offers a rich set of data to challenge state-of-the-art activity classification and detection. This dataset from users with different levels of proficiency, includes foldable non-rigid objects and at various task completion lengths, as shown in Fig. 1 and 2. We also present results on offline, online action classification and action detection as a benchmark for future research.

The EPIC-Tent dataset provides a set of action and error labels, eye-gaze positions and participants' frame-level stage uncertainty self-ratings. It consists of videos and labels from participants assembling a tent outdoors while wearing an eye tracking system (SMI eye tracking glasses, Sensorimotoric Instruments GmbH, Berlin, Germany) and a GoPro HD camera (Go Pro Inc, San Mateo, CA, USA), both head-mounted, as shown in Fig. 1. Participants wore

[1]The EPIC-Tent dataset is available from authors' webpages.

the cameras outdoors in a garden on campus, and were told to assemble the tent however they liked using the printed instructions as needed. We collected data from 24 participants, all of whom answered a brief questionnaire on their level of experience. An observer annotated the videos with task/sub-task labels and error labels (Section 3). Participants subsequently provided a frame-by-frame rating of stage uncertainty by watching their performance from the GoPro video, that is they were rating how certain they felt performing each subtask. We consider this procedure better than interfering with the activity as it is happening. Furthermore, we believe this type of uncertainty information is novel to provide and enabling of future studies in egocentric perception. The total duration of recordings for all participants is over 5 hours, with a total number of clips of 24×3 (3 for the following video types: SMI without eye gaze, SMI with eye gaze, and GoPro video). The average frame rate for the GoPro video is 60 Hz, SMI with eye gaze is 30Hz and 24Hz for SMI without eye gaze. The resolution of videos are 1920×1080 for GoPro and 1280×960 for both SMI videos. Eye gaze text data files were saved at 60Hz.

The SMI and GoPro videos were unsynchronised during recording but manually synchronised afterwards. All activity annotations were made using the SMI video. We then converted these annotations to be synchronized with both the GoPro and SMI videos. The EPIC-Tent dataset provides timestamps for the start and end for video segments as well as frame indices.

The EPIC-Tent dataset includes a rich set of information as follows. **Task / Subtask labels**, there are 38 individual subtasks which correspond hierarchically to 12 simplified tasks that represent the main events required to build a tent, as shown in Table 1. **Error labels**, there are 8 error labels that occur when participants build a tent according to their personal understanding or proficiency. **Uncertainty ratio**, each frame of the GoPro video has a self-evaluated uncertainty rating between 0 (low uncertainty) and 1 (highly uncertain) provided by their respective participants. **Eye gaze position** containing the 2D coordinates of where a participant is looking in the SMI video. **Questionnaire responses**, each participant answered five questions about their level of experience.

The challenges of the EPIC-Tent dataset are as follows. **Overlapped actions**, a task (e.g., instruction reading) can interrupt other tasks. Identifying or recognising these segments is a challenging problem due to the small set of examples. **Non-rigid objects**, the tent, bags, and support bars are non-rigid and deform in a variety of ways presenting a visual recognition challenge. **Varied Task Durations**, tasks within the same class can vary in duration by 10 of seconds both within and between participants. **Egocentric motion**, ego-motion is a natural problem when capturing an activity using a wearable camera. **Coexistence of important objects**, the coexistence of multiple task relevant objects in the same scene can confuse predictions.

The remainder of the paper is organised as follows: In Section 2, we review related work in egocentric datasets, action recognition, and uncertainty prediction. In Section 3, we describe the data annotation method. Benchmarks and baseline results using the proposed EPIC-Tent dataset are provided in Section 4. Conclusions are presented in Section 5.

2. Related Work

Egocentric dataset: Computer vision research on egocentric viewpoints falls into several topics. The most common is focused on understanding activities in a particular environment, such as a kitchen [9, 30], a house [36], an office [1] or outdoors [44]. More specific topics include hand-focused topics (e.g., pose [4], gesture [52], interaction with objects [14, 33, 40, 28]), situation-relevant object-related topics (e.g., food [8], daily life [7, 39], task [10]), social activity-relevant topics (e.g., social pattern [2], social interaction [49, 13], search [50] and summary [27, 31] in daily life videos), and contextual activity-relevant topics (e.g., multi-modal data [41, 35], events [37], life log [6, 48, 12, 43]). We note these studies have not addressed recognition of specific task-related activities that could include errors and uncertainties that change task completion times.

Action recognition: Egocentric video has unique characteristics, but state-of-the-art deep learning methods are still applicable to action recognition [54, 11] and localisation [22, 42] using RGB videos in egocentric viewpoint [29]. However, fast processing times and recognition accuracy are critical factors in providing real-time feedback in an egocentric system. Among state-of-the-art online action recognition [54, 51, 19, 23, 26] and localisation [42, 45] methods, ECO [54] provides good performance in terms of both processing time and action recognition accuracy. We here use the ECO network [54] to provide offline and online action recognition baselines for the EPIC-Tent dataset. Action detection was not addressed by the original ECO network authors, yet we introduce a simple method to detect actions utilising frame-level predictions from the online action recognition model.

Uncertainty prediction: Uncertainty and the related concept of confidence have been frequently studied within psychology within the domain of decision making and perception [38, 25, 5]. [32] examined action uncertainty, demonstrating when participants choose between difficult choices their computer mouse trajectories are altered. During novel tool use [34, 20] found that in actors and observers, respectively, prior knowledge and uncertainty alter problem solving. Uncertainty in eye movements has been

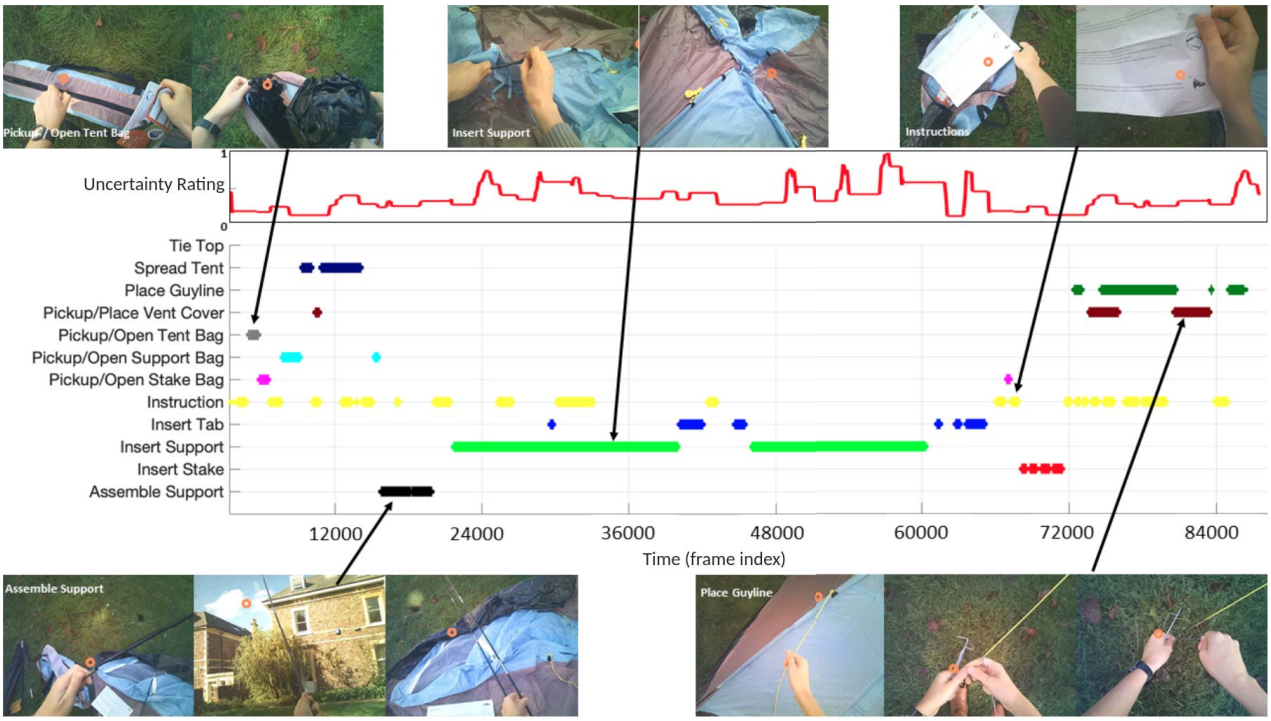


Figure 2. Example time line of activities and uncertainty rating for a single participant: Screen captures illustrate video data from the SMI video with eye gaze cursor for several example action classes.

examined both during cognitive decisions [15], problem solving while copying a model [3], outdoor navigation [17] and driving [46]. However, to our knowledge there have been no prior attempts to combine egocentric video, eye movement data, and frame-by-frame uncertainty ratings for a complex extended task like ours. The EPIC-Tent dataset opens the possibility to combine these psychological measures with state-of-the-art computational methods like deep learning.

3. The EPIC-TENT Dataset

We recorded participants assembling a camping tent outdoors while wearing a head mounted camera and eye tracker. Participants setup a tent (Wilko 2-person dome tent) in a grassy area and were allowed to take as much time as they required and use the printed instructions as desired. 24 participants (14 female, 10 male, mean age 23.3 and std.dev. 4.8) gave informed consent including authorization for anonymized open data sharing. All participants were informed that data were gathered as part of a project to design a wearable computing system that will deliver real-time assistance to wearers.

3.1. Video Annotations

Video data were annotated by observers to delineate tasks, sub-tasks and errors involved in assembly. For the uncertainty rating, after tent assembly, participants viewed their first-person camera GoPro video and rated their level of task uncertainty frame-by-frame using a specialized video viewer. Gaze position was recorded both on the SMI scene camera video (30 Hz) and in text file (60 Hz).

Subtask Labelling: From watching participants' videos, we established a set of common events to be annotated, see Table 1 for a list of tasks/subtasks and the typical workflow in building a tent. Subtask divisions were somewhat arbitrary, but motivated by the pragmatics of manual coding. As some subtasks are equivalent, for instance staking a tent corner is coded as a unique event referencing a particular corner, but can be considered a repetition of a generic corner staking sub-task. We were able to compress setting up the tent into 12 generic labels along with a background label used when the participant was transitioning between subtasks, e.g. walking around the tent. One observer annotated this task information.

Eye Tracking Quality: Eye tracking calibration accuracy was evaluated at the beginning and end of tent assembly. The experimenter moved a fixation target in front of the participant and asked them to fixate the point for 1∼2 sec. An observer recorded the angular distance between the eye gaze cursor and the intended point of fixation, once for each test point. On average, 10 points were tested before and after assembly (this varied across participants as sometimes

Table 1. The details of sub-task annotation

Subtask	Task	Index	Number of samples in task	Number of samples in subtask
Assemble support1	Assemble support	0	56	29
Assemble support2				27
Insert stake c1	Insert stake	1	127	35
Insert stake c2				28
Insert stake c3				36
Insert stake c4				28
Insert support1 c1	Insert support	2	57	4
Insert support1 c2				9
Insert support1 c3				7
Insert support1 c4				9
Insert support2 c1				15
Insert support2 c2				19
Insert support2 c3				21
Insert support2 c4				22
Insert support1 tab c1	Insert support tab	3	147	9
Insert support1 tab c2				9
Insert support1 tab c3				5
Insert support1 tab c4				5
Insert support2 tab c1				17
Insert support2 tab c2				21
Insert support2 tab c3				17
Insert support2 tab c4				15
Instruction	Instruction	4	205	205
Pickup/open stake bag	Pickup/open stake bag	5	53	53
Pickup/open support bag	Pickup/open support bag	6	32	32
Pickup/open tent bag	Pickup/open tent bag	7	26	26
Pickup/place vent cover	Pickup/place vent cover	8	35	35
Place guyline c1	Place guyline	9	126	30
Place guyline c1//c2				4
Place guyline c1//c3				1
Place guyline c2				30
Place guyline c2//c3//c4				1
Place guyline c2//c4				1
Place guyline c3				26
Place guyline c3//c4				4
Place guyline c4				29
Spread tent	Spread tent	10	42	42
Tie top	Tie top	11	15	15

the point was outside the camera field of view). Median tracker accuracy (Euclidean distance between the target and point of gaze cursor) was calculated across the points per individual, and averaged across participants, yielding an average of $1.87^\circ \pm 1.16^\circ$ (range: $0.68^\circ \sim 4.48^\circ$). Three participants $(1, 13, 21)$ have high error $> 5^\circ$ and are recommended to be excluded in eye tracking coordinate analysis.

Uncertainty Labelling: Participants viewed and rated their performance by watching the GoPro video using a Matlab video viewer that let the participant control the video (forward, backwards, speed up and down) and give a continuous uncertainty rating using mouse position. Participants gave a rating between 0 (low uncertainty) and 1 (highly uncertain) for each frame of the video. To help motivate giving good self-ratings, before rating the video, participants were told of our goal to build a digital assistant and to imagine when they would have required help. Note, only particpants 8-24 gave these ratings.

Error Labelling: Two observers annotated the eye tracking videos for errors including: 1) Motor errors, 2) misuse of equipment, 3) steps out of order, 4) equipment failure, 5) omission of a step, 6) searching for an item, 7) correction of a prior error, 8) slowness in movement, and 9) repetition. All errors were annotated with begin and end times, except omission. Ordering errors were marked for the duration of the out of order subtasks, where the correct order was defined as the sequence in the printed instructions.

Statistics of videos: EPIC-Tent dataset involves videos over 5.4 hours. There are 24 videos containing individ-

ual subject in each video. The total number of frames in the dataset is $1,171,897$. The average length of the video is 13.6 minutes. The average total number of frames is $48,829$. The total number of action instances in the dataset is 921.

3.2. Participant Behaviour

To assess participants' experience, they completed a survey asking: (1) How often do you camp each year?; (2) Are you an experienced camper?; (3) Do you own a tent?; (4) How many times a year do you setup a tent?; (5) How many times have you ever setup a tent?

Questionnaire responses were turned into a numeric score and normalized by dividing by the maximum response across participants, and then averaged to yield a number between 0 and 1 that represented past experience with camping and tents. Participants generally had a low amount of prior experience setting up a tent. Mean self-rating of experience, 1.4 out of 5, (std.dev. 1.3, range: $0 \sim 4.1$). Self-rated experience was negatively correlated with total time spent reading the instructions r(22)=-0.56, p=0.005, β=-36 sec., intercept=151 sec., i.e. the more self-rated experience, the less time spent reading instructions. Participants took between 4.3 to 24.1 minutes to complete the task (mean=12.1 min., median=11, std.dev.=4.7). All participants were able to erect the tent, albeit sometimes with slight problems, e.g. forgetting the vent cover, not tying the support beams to the top of the tent or staking the guy lines incorrectly. Note, self-rated experience did not correlate significantly with duration, r(22)=-0.31, p=0.14.

4. Benchmarks and Baseline Results

The EPIC-Tent dataset lends itself to a range of possible challenges, including predicting uncertainty, recognising errors and omissions, predicting the level of expertise, as well as predicting and anticipating gaze/attention. We leave such challenges to future research and focus on three standard benchmarks.

We provide baseline results on the EPIC-Tent dataset for three tasks: offline action recognition (Off-AR), online action recognition (On-AR), and online action detection (On-AD). We use a state-of-the-art action recognition architecture for video (ECO network) [54]. The ECO network is light-weight, able to process long videos while maintaining real-time performance, and produces competitive results on several action recognition datasets (tested on UCF101, HMDB, Kinetics and Something-Something). The ECO architecture uses BN-Inception as a backbone [47] for the 2D convolutional layers and several layers of 3D-Resnet-18 [18] for the 3D convolutional layers to capture the spatial and temporal features, respectively. Results in this section use $ECO_{Lite-16F}$. We also use two pre-training set of weights, one on Kinetics [24] and the other on Something-Something V2 [16] - both weights provided in the ECO github repo. We report the best result from either initialisation in each case.

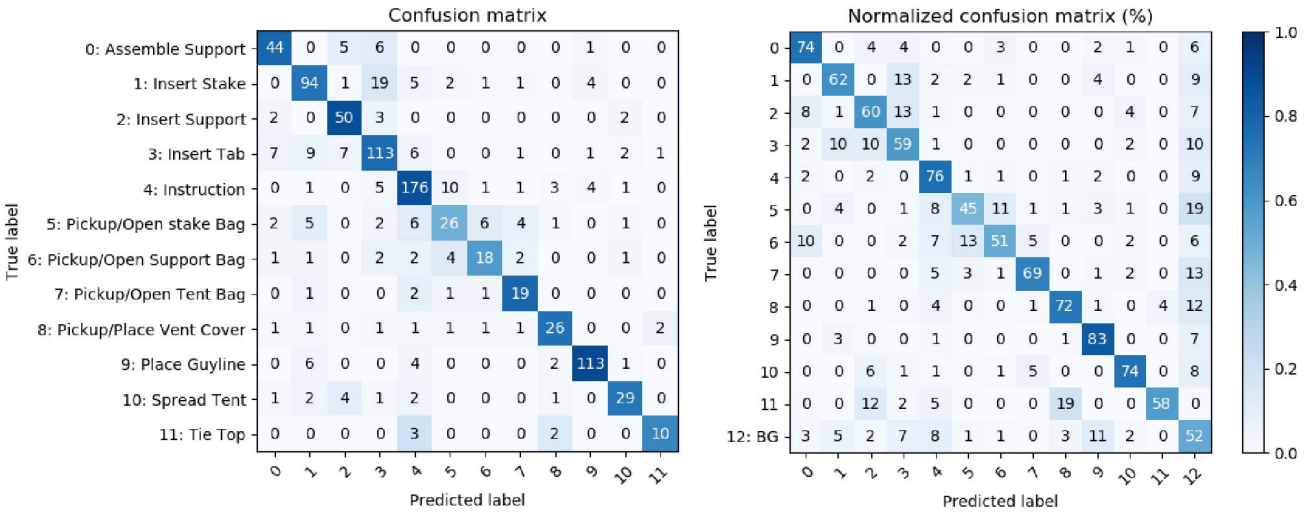
We followed most of the detailed implementations used to train the original ECO architecture. We resized the resolution of 1920×1080 of the original GoPro image to 456×256 and employed fixed-corner cropping and scale jittering with horizontal flipping, per-pixel mean subtraction and resized the cropped regions to 224×224, as in [54]. We used a mini-batch size of 12, and randomly selected one frame from each evenly split 16 segments. For Off-AR, learning rate was initialised to 0.01 decaying by a factor of 10 every 50 epochs. The learning was halted after 200 epochs. For On-AR/AD, the learning rate decayed every 25 epochs and the learning was halted after 100 epochs. We fine-tuned the network with a momentum of 0.9, a weight decay of 0.0005. Dropout ratio is 0.3.

For training and testing, we use 4-fold cross validation. We shuffle the videos then split the dataset, at the video level, so a full sequence is either in training or in testing. For each fold, one quarter of the dataset is used for testing with the remaining 3/4's of the dataset for training. We accumulate the results of all folds to report the performance on the whole dataset for each task. The folds will be released alongside the dataset annotations, to allow replicating the results.

For each fold, the whole fine-tuning process on the proposed EPIC-Tent took around three hours and ten hours, in Off-AR and On-AR respectively, on two GeForce GTX 1080 Ti. On the hardware, inference was achieved at 186 video clips per second (VPS) with 5.37 ms average processing time.

4.1. Offline Action Recognition Benchmark

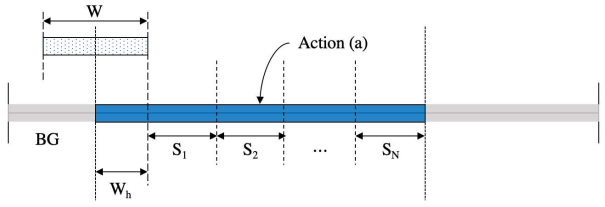
In offline action recognition (Off-AR), we use the labeled start/end times of each action, and train to classify the video segment as one of 12 task labels (see Table 1). We do not learn or predict the background class similar to customary Off-AR approaches that assume the extents of relevant actions have been predefined. We report results in the confusion matrix in Fig. 3(a), using the normalised performance in the colourmap, but also report the number of test segments in each confusion cell. The overall accuracy for Off-AR is **78.64**%. Specifically, the first, second, third, and fourth fold accuracies are 74.43%, 83.57%, 79.90%, and 77.78%, respectively. These results use Something-Something V2 pre-training weight, which mariginally outperformed pre-training on Kinetics (**77.66**%). The largest confusion is observed in temporally neighbouring and overlapping tasks, particularly when the number of training samples is relatively small.



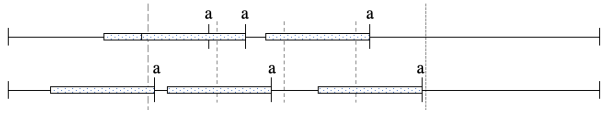
(a) Offline Action Recognition (Accuracy: 78.64%)

(b) Online Action Recognition (Accuracy: 64.08%)

Figure 3. Confusion matrices for offline ((a)) and online ((b)) action recognition using four-fold cross-validation: (a) The number in a cell represents the number of predicted actions. (b) The number in a cell represents the normalised accuracy (percent) of each class.



(a) Segmenting actions in On-AR



(b) Samples for training

Figure 4. Examples of dataset sampling for online action recognition: (a) segmenting an action into N segments and (b) sampled windows from an action annotation during training.

4.2. Online Action Recognition Benchmark

The goal of online action recognition (On-AR) is to recognise the ongoing task, given the current frame and preceding frames (i.e., history) of a fixed or variable length. We report our baseline On-AR results using a fixed window length W which we set to 350 frames. When actions overlap, we use the shorter action as the online frame ground truth label. This is because shorter actions in fact interrupt longer ones (e.g. checking instructions while spreading the tent).

During On-AR, both task-relevant and background tasks would take place. Thus, we add an additional class to represent background frames. As the dataset has actions of variable length, we avoid oversampling windows of longer actions during training, by sampling the same number of windows from every annotated action, including the background segments. We show our sampling strategy in Fig. 4. We divide each annotated action uniformly into N segments $S_i = \{S_1, \cdots, S_N\}$, after subtracting $W_h = \frac{W}{2}$ frames from the start. This ensures at least half the window contains the relevant action, similar to the approach used in [53] for online action recognition. In this paper, N is 5. Then,we randomly sample one frame within the each segment, as the observed frame, along with a history of size W. Fig. 4(b) shows how multiple windows would be extracted over different iterations in training. This offers a natural data augmentation strategy, while ensuring 1) actions of various lengths are equally represented, and 2) the various stages of the task from start to conclusion are included in training.

In testing, a sliding window of the same width and a stride of 22 frames is utilised, sampling 16 frames uniformly in each window. We report On-AR results over the test set, using the same 4-folds in training/testing as in Sec 4.1. Experimental results show that the overall accuracy of On-AR is 64.08%, as shown in Fig. 3(b). Specifically, the first, second, third, and fourth fold accuracies are 66.37%, 68.01%, 62.12%, and 58.75%, respectively. These results use pre-training weights from Kinetics.

Qualitative results are shown in Fig 5, comparing the ground truth to On-AR predictions on two parts of the same video. Errors around frame 8000 show the complexity of recognising tasks such as 'inserting support' and 'assembling support'.

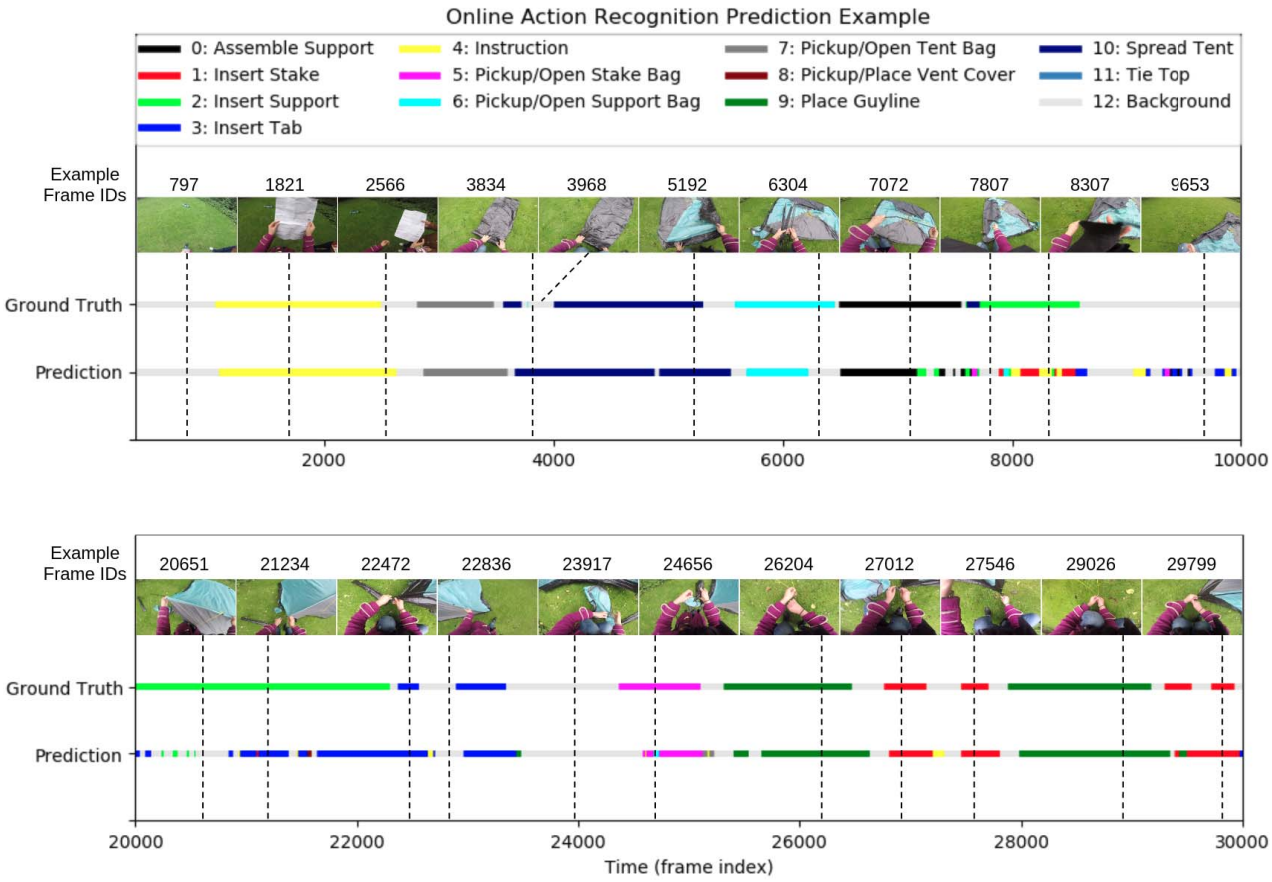


Figure 5. Example of Online Action Recognition (On-AR)

Table 2. Online Action Detection (On-AD) results – allowing overlapped action detection

IoU	mAP	Average Precision (AP) for Each Action											
		0	1	2	3	4	5	6	7	8	9	10	11
0.1	68.65	73.83	70.20	71.51	67.23	74.68	40.05	56.07	78.65	74.35	80.92	73.31	63.02
0.2	69.85	80.78	70.71	67.09	64.16	78.70	32.50	57.79	78.83	72.43	86.20	80.09	68.96
0.3	70.41	81.50	67.26	62.46	64.53	80.02	36.78	54.43	78.83	71.93	88.12	86.11	72.95
0.4	67.10	76.09	57.63	58.84	64.26	80.52	36.79	51.08	75.92	64.23	83.62	87.58	68.67
0.5	**64.31**	74.87	51.52	56.04	57.86	76.03	36.19	50.78	73.03	67.92	70.54	85.12	71.83
0.6	57.30	65.95	41.25	48.63	54.93	70.36	37.55	49.80	58.09	51.94	60.58	76.73	71.83
0.7	51.88	65.42	29.89	44.77	48.57	64.51	33.28	36.17	42.32	51.89	55.97	73.39	76.36
0.8	37.94	33.77	17.72	43.24	26.99	58.40	09.26	29.86	27.78	51.04	44.36	46.53	66.27

4.3. Online Action Detection Benchmark

We report results for Online Action Detection (On-AD), where the task is to localise the extent of actions, including overlapping actions/tasks. We use the same train/test splits and model as in Sec. 4.2. For each class, we consider all consecutive frames with class confidence above a threshold $\alpha = \{0.01, 0.02, \cdots, 1\}$. This is used to plot the Precision-Recall curve, from which we calculate the interpolated average precision for each class, as in [21]. We report average precision as well as class mean average precision (mAP) for various IoU = $\{0.1, 0.2, \cdots, 0.8\}$ in Table. 2. At IoU 0.5, we report an overall mAP of **64.31%**. From the table, we note that tasks involving inserting stake/support (1-2) and opening stake/support bags (5-6) are the hardest to detect.

5. Conclusion and Future Work

In this paper we have described the process of collecting, annotating and benchmarking an egocentric video dataset of natural behavior while participants assembled a camping tent (EPIC-Tent). The dataset features natural outdoor

lighting, behavior from novice to moderately experienced participants, and a wide variance in examples of each action class. It offers a rich set of data from egocentric video, eye-tracking data, self-rated uncertainty for each video frame, and error in performance labels. In future work, we plan to use convolutional networks to relate visual features to the uncertainty rating and error labels. The EPIC-Tent dataset opens the possibility to combine these psychological measures with state-of-the-art computational methods to help push forward research in egocentric perception.

Acknowledgments

Access to the EPIC-Tent dataset and annotations available from authors' webpages. Supported by UK EPSRC GLANCE (EP/N013964/1).

References

[1] G. Abebe, A. Catala, and A. Cavallaro. A first-person vision dataset of office activities. In *Proc. of Int. Workshop on Multimodal Pattern Recognition of Social Signals in Human Computer Interaction*, Beijing, China, 2018.

[2] M. Aghaei, M. Dimiccoli, C. C. Ferrer, and P. Radeva. Towards social pattern characterization in egocentric photo-streams. *Computer Vision and Image Understanding*, 2018.

[3] D. H. Ballard, M. M. Hayhoe, and J. B. Pelz. Memory representations in natural tasks. *Journal of cognitive neuroscience*, 7(1):66–80, 1995.

[4] S. Bambach, S. Lee, D. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[5] A. R. Bland et al. Different varieties of uncertainty in human decision-making. *Frontiers in neuroscience*, 6:85, 2012.

[6] M. Bolaños, Á. Peris, F. Casacuberta, S. Soler, and P. Radeva. Egocentric video description based on temporally-linked sequences. *J. Visual Communication and Image Representation*, 50:205–216, 2018.

[7] M. Bolaños and P. Radeva. Ego-object discovery. *CoRR*, abs/1504.01639, 2015.

[8] M. Bolaños and P. Radeva. Simultaneous food localization and recognition. In *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, pages 3140–3145, 2016.

[9] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[10] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *Proceedings of the British Machine Vision Conference*, 2014.

[11] A. Diba, M. Fayyaz, V. Sharma, M. M. Arzani, R. Yousefzadeh, J. Gall, and L. V. Gool. Spatio-temporal channel correlation networks for action classification. In *ECCV*, 2018.

[12] M. Dimiccoli, M. Bolaos, E. Talavera, M. Aghaei, S. G. Nikolov, and P. Radeva. Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation. *Computer Vision and Image Understanding*, 155:55 – 69, 2017.

[13] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233, June 2012.

[14] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2018.

[15] K. Gidlöf, A. Wallin, R. Dewhurst, and K. Holmqvist. Using eye tracking to trace a cognitive process: Gaze behaviour during decision making in a natural environment. 2013.

[16] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fründ, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5843–5851, 2017.

[17] M. M. Hayhoe and J. S. Matthis. Control of gaze in natural environments: effects of rewards and costs, uncertainty and memory in target selection. *Interface focus*, 8(4), 2018.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[19] B. Hu, J. Yuan, and Y. Wu. Discriminative action states discovery for online action recognition. *IEEE Signal Processing Letters*, 23(10):1374–1378, Oct 2016.

[20] P. O. Jacquet, V. Chambon, A. M. Borghi, and A. Tessari. Object affordances tune observers' prior expectations about tool-use behaviors. *PloS one*, 7(6):e39629, 2012.

[21] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://crcv.ucf.edu/THUMOS14/`, 2014.

[22] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid. Action tubelet detector for spatio-temporal action localization. In *International Conference on Computer Vision (ICCV 2017)*, pages 4415–4423. Institute of Electrical and Electronics Engineers (IEEE), 12 2017.

[23] V. Kantorov and I. Laptev. Efficient feature extraction, encoding, and classification for action recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2593–2600, June 2014.

[24] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[25] A. Kepecs and Z. F. Mainen. A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1322–1337, 2012.

[26] I. Kviatkovsky, E. Rivlin, and I. Shimshoni. Online action recognition using covariance of shape and motion. *Computer Vision and Image Understanding*, 129:15 – 26, 2014.

[27] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1346–1353, June 2012.

[28] C. Li and K. M. Kitani. Pixel-level hand detection in egocentric videos. *CVPR*, 2013.

[29] H. Li, Y. Cai, and W.-S. Zheng. Deep dual relation modeling for egocentric interaction recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[30] Y. Li, M. Liu, and J. M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[31] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2721, June 2013.

[32] C. McKinstry, R. Dale, and M. J. Spivey. Action dynamics reveal parallel competition in decision making. *Psychological Science*, 19(1):22–24, 2008.

[33] 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 *Proceedings of International Conference on Computer Vision (ICCV)*, October 2017.

[34] F. Munoz-Rubke, D. Olson, R. Will, and K. H. James. Functional fixedness in tool use: Learning modality, limitations and individual differences. *Acta psychologica*, 190:11–26, 2018.

[35] K. Nakamura, S. Yeung, A. Alahi, and L. Fei-Fei. Jointly learning energy expenditures and activities using egocentric multimodal signals. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

[37] Y. Poleg, A. Ephrat, S. Peleg, and C. Arora. Compact cnn for indexing egocentric videos. In *WACV*, 2016.

[38] A. Pouget, J. Drugowitsch, and A. Kepecs. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience*, 19(3):366–374, 2016.

[39] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3137–3144, June 2010.

[40] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 2730–2737. IEEE Computer Society, 2013.

[41] M. M. Silva, W. L. S. Ramos, J. P. K. Ferreira, F. C. Chamone, M. F. M. Campos, and E. R. Nascimento. A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2383–2392, Salt Lake City, USA, Jun. 2018.

[42] G. Singh, S. Saha, M. Sapienza, P. H. S. Torr, and F. Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *ICCV*, pages 3657–3666. IEEE Computer Society, 2017.

[43] K. K. Singh, K. Fatahalian, and A. A. Efros. Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016.

[44] S. Singh, C. Arora, and C. Jawahar. Trajectory aligned features for first person action recognition. *Pattern Recognition*, 62:45–55, 2017.

[45] K. Soomro, H. Idrees, and M. Shah. Predicting the where and what of actors and actions through online action localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2648–2657, June 2016.

[46] B. T. Sullivan, L. Johnson, C. A. Rothkopf, D. Ballard, and M. Hayhoe. The role of uncertainty and reward on eye movements in a virtual driving task. *Journal of vision*, 12(13):19–19, 2012.

[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[48] Y. Tang, Y. Tian, J. Lu, J. Feng, and J. Zhou. Action recognition in rgb-d egocentric videos. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3410–3414, Sep. 2017.

[49] R. Yonetani, K. M. Kitani, and Y. Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2629–2638, June 2016.

[50] R. Yonetani, K. M.Kitani, and Y. Sato. Ego-surfing first-person videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, June 2015.

[51] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang. Real-time action recognition with enhanced motion vector cnns. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2718–2726, June 2016.

[52] Y. Zhang, C. Cao, J. Cheng, and H. Lu. EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia*, 20(5):1038–1050, May 2018.

[53] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2923–2932, Oct 2017.

[54] M. Zolfaghari, K. Singh, and T. Brox. ECO: efficient convolutional network for online video understanding. In *ECCV*, 2018.