Supplementary
Toyota Smarthome: Real-World Activities of Daily Living

Srijan Das[1,2], Rui Dai[1,2], Michal Koperski[1,2], Luca Minciullo[3], Lorenzo Garattoni[3], Francois Bremond[1,2] and Gianpiero Francesca[3]

[1]Université Côte d'Azur [2]Inria [3]Toyota Motor Europe

In this supplementary material, we provide more details on the Toyota Smarthome dataset (hereafter Smarthome), the evaluation protocols and the state-of-the-art methods used for comparison in the main body of the paper.

In fig. 1, we present an example frame for each of the 31 activities in Smarthome. The figure shows the rich diversity of activity classes in the dataset: some activities are a composition of sub-activities (e.g., *cooking* is composed of *cleaning dishes, cleaning up, cutting, stirring* and *using stove*), some activities correspond to the same activity, but performed using different objects (e.g., *drinking* from a *cup*, *can*, or *bottle*), other activities are almost completely static (e.g., *reading book*, *using phone* or *watching TV*). To learn more about Toyota Smarthome dataset please visit the project website[1].

To date, there are more than 50 human activity recognition datasets. Although each one of them has unique, beneficial characteristics for the evaluation of activity recognition algorithms, they have also limitations. Table 1 lists the most popular public RGB+D ADL datasets to our knowledge with their key features. All the datasets mentioned in table 1 are captured indoors using Kinect sensors and thus provides either skeleton info or depth map in addition to RGB cue. In terms of dataset size (i.e., number of video samples and activity classes), Smarthome is the second largest dataset with 16,115 clips.

For evaluation of activity recognition algorithms on Smarthome, we defined a cross-subject (CS) and two cross-view (CV_1 & CV_2) protocols. Fig. 2, 3 and 4 depict training and testing sample distributions of activities for the different evaluation protocols. Fig. 2 shows how the activity categories are imbalanced throughout the dataset and in our evaluation protocol. For example, the number of training samples for the activities *walking, drinking from glass, reading book* are considerably higher than with other classes. This feature is another key difference between Smarthome and existing datasets.

We evaluate the cross-view protocol on 19 activity categories, namely: *Cut bread*, *Drink From bottle*, *Drink From can*, *Drink From cup*, *Drink From glass*, *Eat at table*, *Eat Snack*, *Enter*, *Getup*, *Leave*, *Pour From bottle*, *Pour From can*, *Read book*, *Sit down*, *Take pills*, *Use laptop*, *Use tablet*, *Use telephone* and *Walk*. The CV_1 protocol is proposed to test the cross-view activity classification performance in the same scene (i.e., *dining room*). Fig. 3 shows the per-class video sample distribution for CV_1. Selecting only two cameras significantly reduced the amount of training samples, making this protocol highly challenging. For this reason, in the CV_2 protocol we further increase the number of training samples, by adding samples from other cameras. In fig. 4, we can see the increased number of samples for these 19 activities. For instance, the number of *Drinking from glass* instances increased by 46 units.

In table 2, we provide an overview of our hyperparameter selection for the state-of-the-art methods benchmarked on Smarthome. This is to enable reproducibility of the results reported in the paper. Note that the kernel and activity regularizers for I3D and I3D+NL are applied in the *softmax* layer. For I3D+NL, we experimented with various numbers of NL blocks at early and late stages. We obtained the highest accuracy with 1 NL block at the last stage.

To preserve anonymity, we blurred the face of all subjects in the dataset. We quantitatively evaluated that this operation reduces activity classification accuracy by less than 1% in all methods reported in the paper.

We present the confusion matrix for each of the evaluation protocols (CS, CSV_1 and CV_2) in Fig. 5, 6 and 7 respectively. The confusion matrices show that our method can recognize with rather high accuracy even activities that are under-represented in terms of training samples. Take, as an example, the accuracy of 84% that the method achieves on the activity *Pour from kettle*, which is represented only by 79 training samples in the cross-subject pro-

[1] https://project.inria.fr/toyotasmarthome

Table 1. Comparison between Smarthome dataset and some of the other daily living activity datasets for activity recognition. Our dataset is the second largest dataset in daily living activity dataset in terms of activity classes and number of video samples.

Dataset Name	#Subjects	#Activity Class	#Videos	#Viewpoint	Modalities	#Year
CAD-60 [7]	4	12	60	1	RGB+D+Skeleton	2011
RGBD-HuDaAct [4]	30	13	1189	1	RGB+D	2011
MSRDailyActivity3D[8]	10	16	320	1	RGB+D+Skeleton	2012
Act4[2]	24	14	6844	4	RGB+D	2012
CAD-120 [3]	4	10+10	120	1	RGB+D+Skeleton	2013
DML-SmartAction[1]	16	12	932	2	RGB+D	2013
NUCLA[9]	10	10	1475	3	RGB+D+Skeleton	2014
Office Activity[10]	10	20	1180	3	RGB+D	2014
UWA3D Multiview II[5]	10	30	1075	5	RGB+D+Skeleton	2015
NTU RGB+D [6]	40	60	56880	80	RGB+D+IR+Skeleton	2016
Smarthome	18	31	16129	7	RGB+D+Skeleton	2019

Table 2. Hyperparameter specifications for various state-of-the-art methods validated on Smarthome.

Methods	**Hyper-parameter**	*CS*	CV_1	CV_2
LRCN	# Neurons	256	128	128
	Gradient clipping	1	1	1
	Dropout	0.5	0.6	0.5
LSTM	# Neurons	512	128	128
	Gradient clipping	1	1	1
	Dropout	0.5	0.6	0.5
I3D	Kernel Regularization	L_2 (0.01)	L_2 (0.01)	L_2 (0.01)
	Actitvity Regularization	L_1 (0.01)	L_1 (0.01)	L_1 (0.01)
	Dropout	0.2	0.5	0.5
I3D+NL	# NL blocks	1	1	1
	Kernel Regularization	L_2 (0.01)	L_2 (0.01)	L_2 (0.01)
	Activity Regularization	L_1 (0.01)	L_1 (0.01)	L_1 (0.01)
	Dropout	0.2	0.5	0.5

tocol (see Fig. 2). The absence of a strong bias towards over-represented classes is confirmed by the mean per-class accuracy: 54.2% for CS, 35.2% for CV_1 and 50.3% for CV_2.

In Fig. 8, we show the spatial and temporal attention masks quantitatively in the form of heatmaps on sample Smarthome videos. The key images (in 1^{st} column) represents an image sampled from the video segment with maximum temporal attention score (segment marked with bright color in 3^{rd} column). The spatial heatmaps (in 2^{nd} column) shows that the attention scores are higher where the action is localized (depicted with green bounding box). This enables the proposed spatio-temporal attention mechanism to better encode the appearance as compared to the baseline I3D network.

References

[1] S. Mohsen Amiri, Mahsa T. Pourazad, Panos Nasiopoulos, and Victor C.M. Leung. Non-intrusive human activity monitoring in a smart home environment. In *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2013.

[2] Zhongwei Cheng, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian. Human daily action analysis with multi-view and color-depth data. In *European Conference on Computer Vision(ECCV)*, 2012.

[3] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. In *IJRR*, 2013.

[4] Bingbing Ni, Gang Wang, and Pierre Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Nov 2011.

[5] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for cross-view action recognition. In *TPAMI*, 2016.

[6] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analy-

sis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[7] Jaeyong Sung, and Bart Selman Colin Ponce, and Ashutosh Saxena. Human activity detection from rgbd images. In *AAAI workshop*, 2011.

[8] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[9] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning, and recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2649–2656, June 2014.

[10] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition and detection by combining motion and appearance features. In *THUMOS*, 2014.

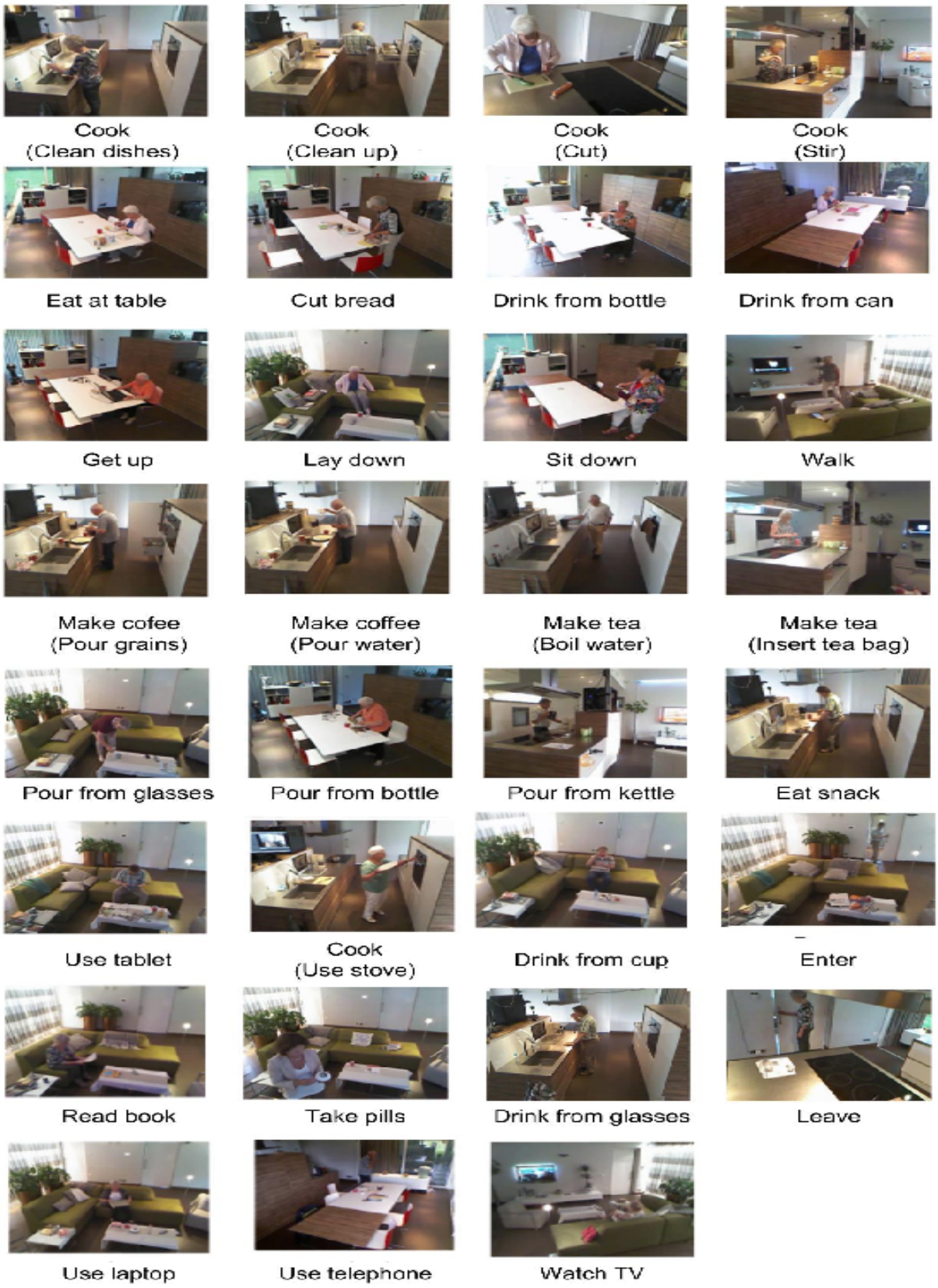

Figure 1. A glimpse of the 31 activity classes in Smarthome.

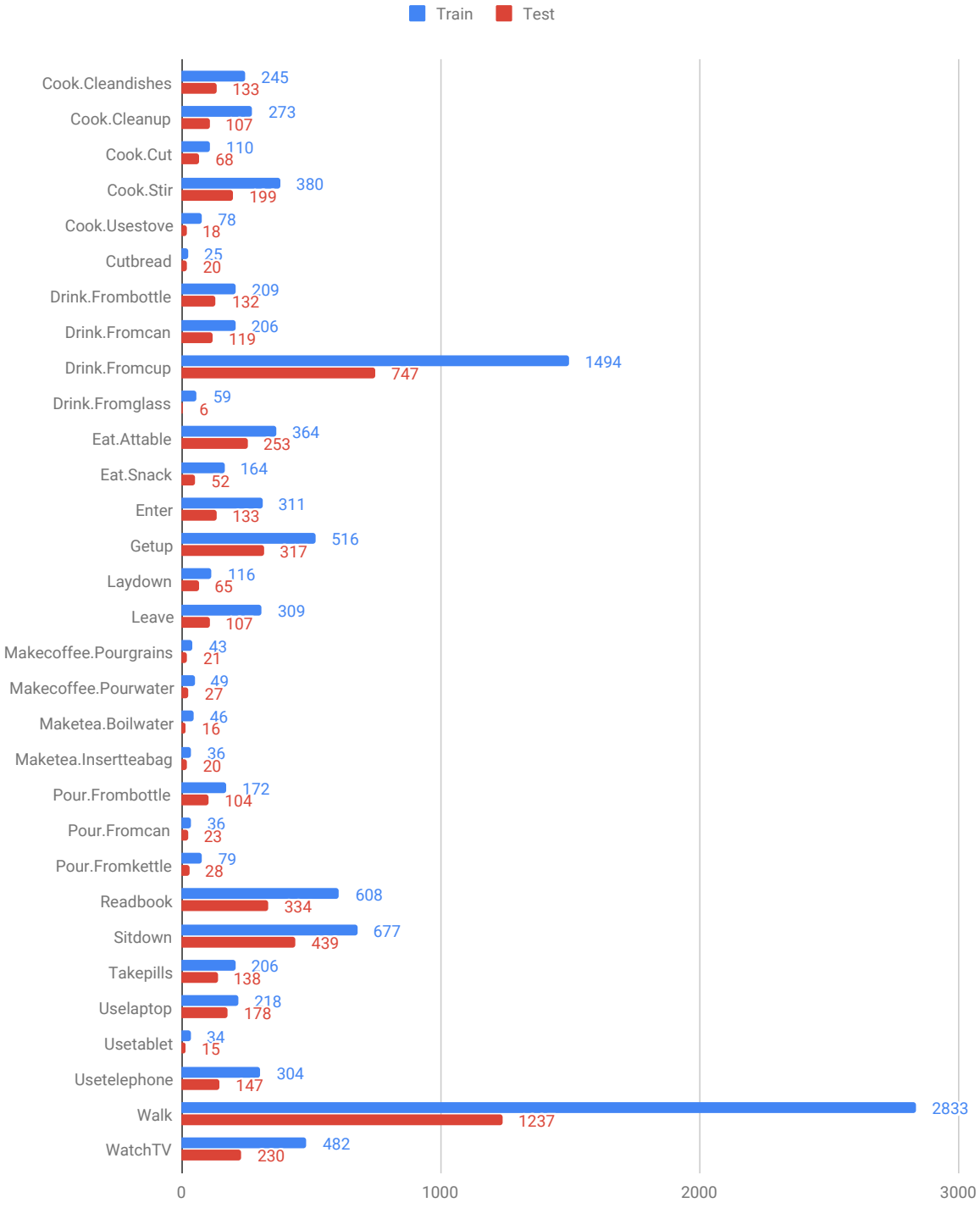

Figure 2. Training/Testing video sample distribution for 31 activity categories in Cross-Subject (CS) protocol.

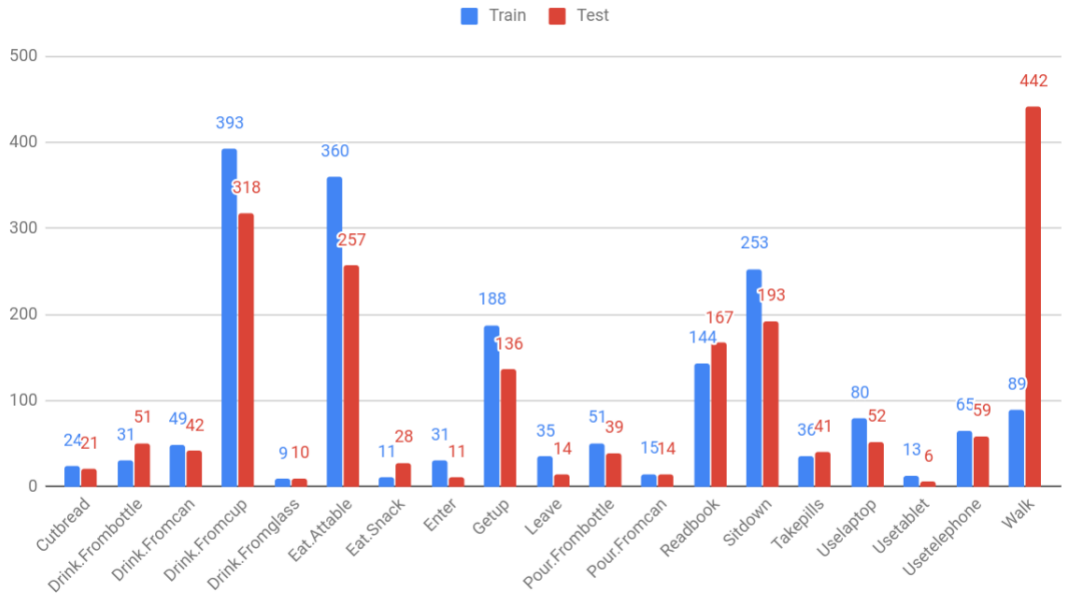

Figure 3. Training/Testing video sample distribution for 19 activity categories in Cross-View1 (CV_1) protocol.

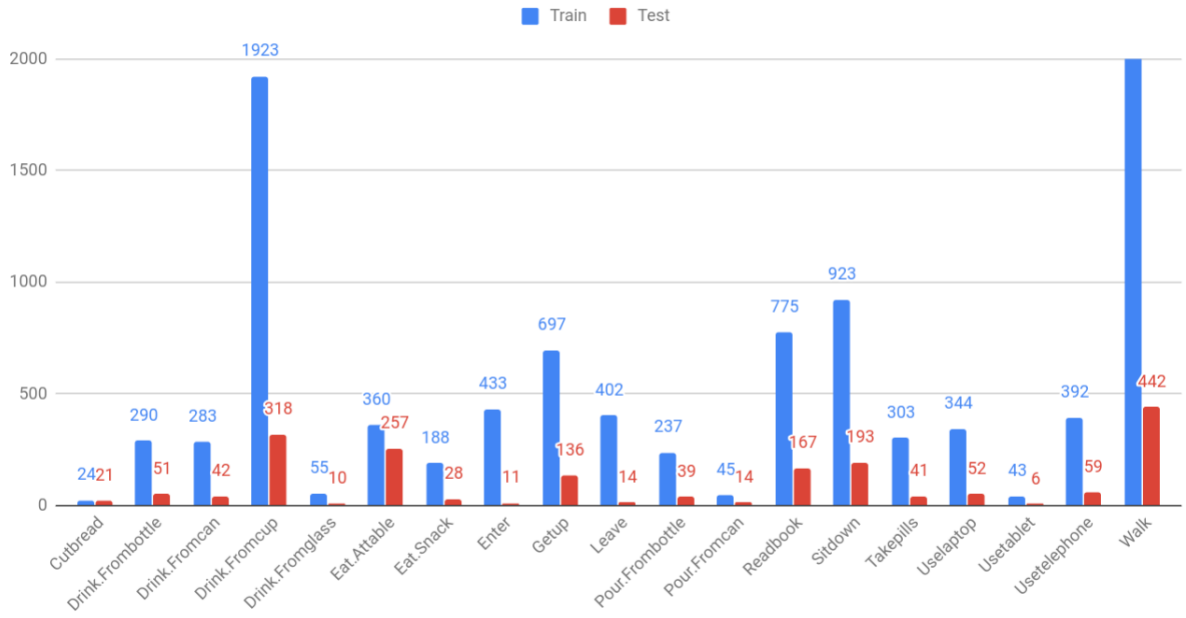

Figure 4. Training/Testing video sample distribution for 19 activity categories in Cross-View2 (CV_2) protocol.

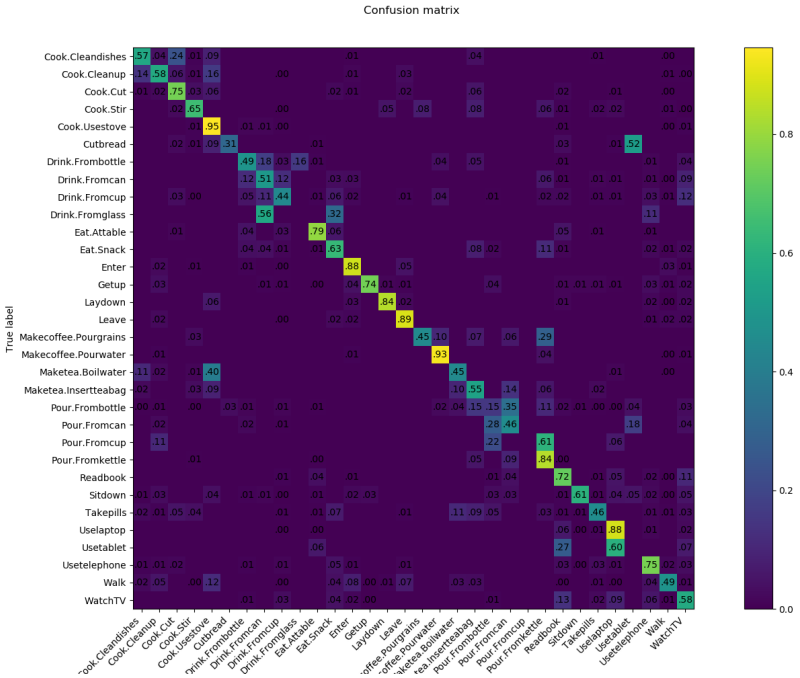

Figure 5. Confusion Matrix for Smarthome in Cross-Subject (CS) protocol using the proposed Separable spatio-temporal attention network. **Note:** We remove the activity label *Pour.Fromcup* due to extremely low number of training samples.

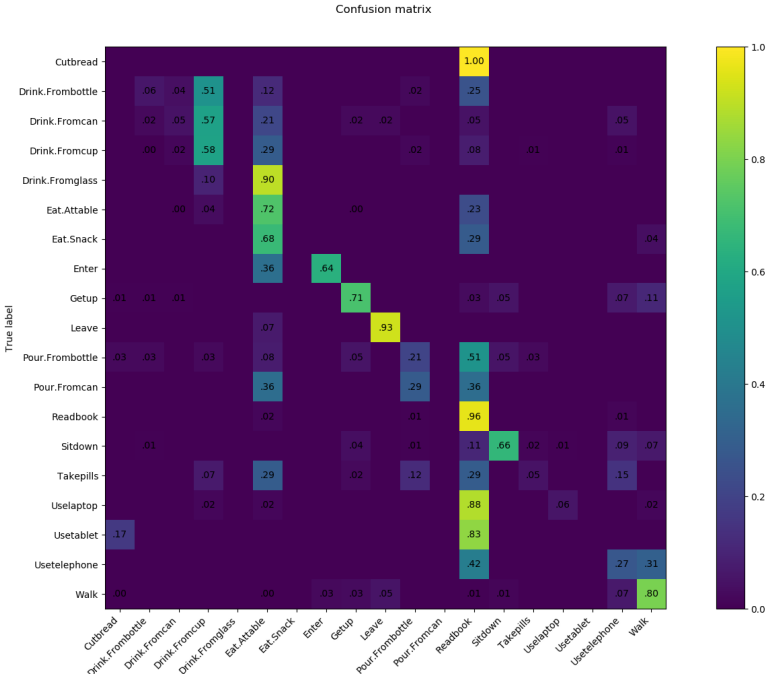

Figure 6. Confusion Matrix for Smarthome in Cross-View 1 (CV_1) protocol using the proposed Separable spatio-temporal attention network.

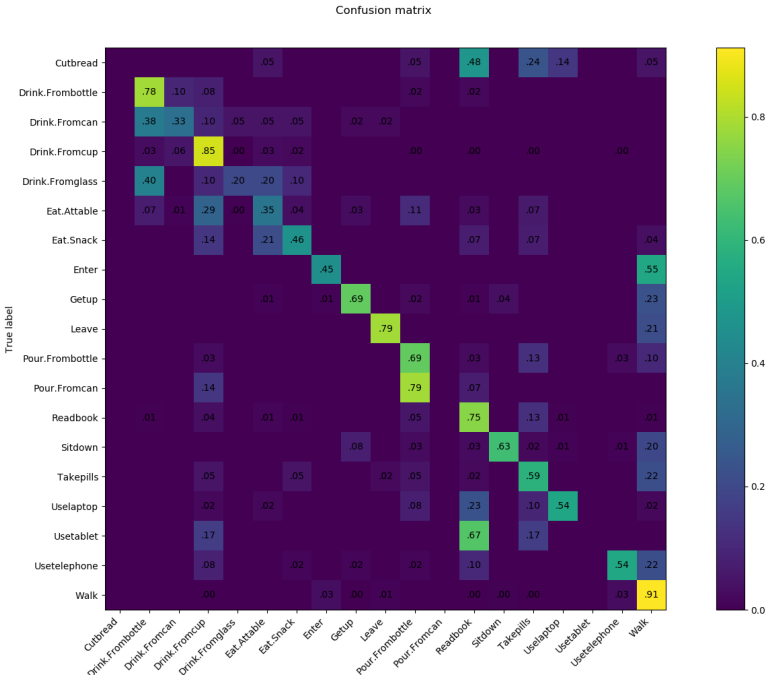

Figure 7. Confusion Matrix for Smarthome in Cross-View 2 (CV_2) protocol using the proposed Separable spatio-temporal attention network.

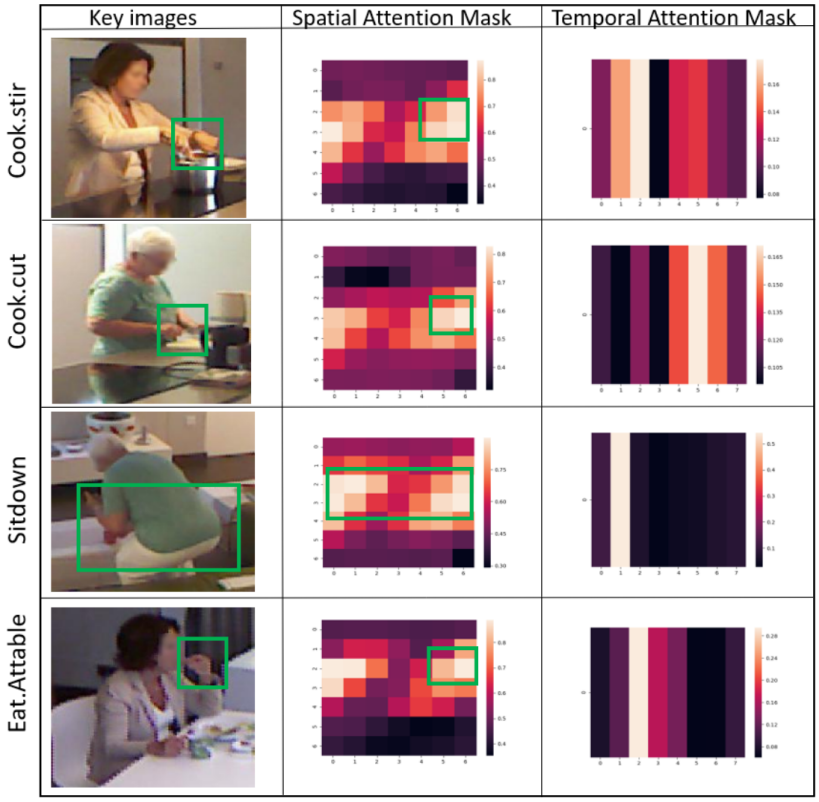

Figure 8. Spatial (2^{nd} column) and Temporal (3^{rd} column) attention mask for Smarthome using the proposed Separable spatio-temporal attention network. The key image (1^{st} column) represents a sample from the segment with high temporal attention score. The green bounding box refers to the action localized coordinates interpolated on the spatial attention mask.