Joint Prediction of Activity Labels and Starting Times in Untrimmed Videos

Supplementary Materials

Tahmida Mahmud[1], Mahmudul Hasan[2], Amit K. Roy-Chowdhury[1]
[1]University of California, Riverside, CA-92521, USA
[2]Comcast Labs, Washington, DC-20005, USA
tmahm001@ucr.edu, mahmud.ucr@gmail.com, amitrc@ee.ucr.edu

Table of Contents

1. Future Activity Label Prediction

1.1. MPII-Cooking Dataset: Activity Class-wise Precision and Recall

Activity Class	Precision	Recall	Activity Class	Precision	Recall
1	0.98	0.99	34	0	0
2	1	0.94	35	0.62	0.86
3	0.53	0.53	36	0.61	0.63
4	0.08	0.06	37	-	0
5	0	0	38	0.80	0.83
6	0.38	0.30	39	0.80	0.83
7	0.96	0.52	40	0	0
8	0.33	0.55	41	0.83	0.92
9	0.25	0.17	42	1	1
10	0.88	1	43	0.5	0.21
11	1	1	44	1	1
12	1	1	45	1	0.87
13	0.50	0.33	46	0.19	0.33
14	0.50	0.40	47	0.95	0.67
15	0.53	1	48	0.25	0.50
16	0.75	0.67	49	0	0
17	0.67	0.50	50	0.64	0.92
18	0.83	1	51	0.67	1
19	1	1	52	0.67	0.28
20	0.97	0.98	53	0.90	0.49
21	1	1	54	0.92	0.85
22	1	1	55	0.91	0.97
23	0.75	0.28	56	0.97	0.85
24	0.61	0.71	57	1	0.33
25	1	1	58	0.59	0.75
26	0.98	0.88	59	0.50	0.33
27	0.83	1	60	0.64	0.85
28	1	1	61	0.83	1
29	0.74	0.63	62	0.93	0.83
30	1	0.90	63	0.66	0.77
31	0.66	1	64	0.67	0.50
32	0.57	0.69	65	0.93	0.58
33	0.67	0.46			

Table 1. Activity class-wise precision and recall for MPII-Cooking Dataset. The missing value corresponds to the activity class which did not occur in the testing sets.

Go to Table of Contents

1.2. MPII-Cooking Dataset: Confusion Matrix

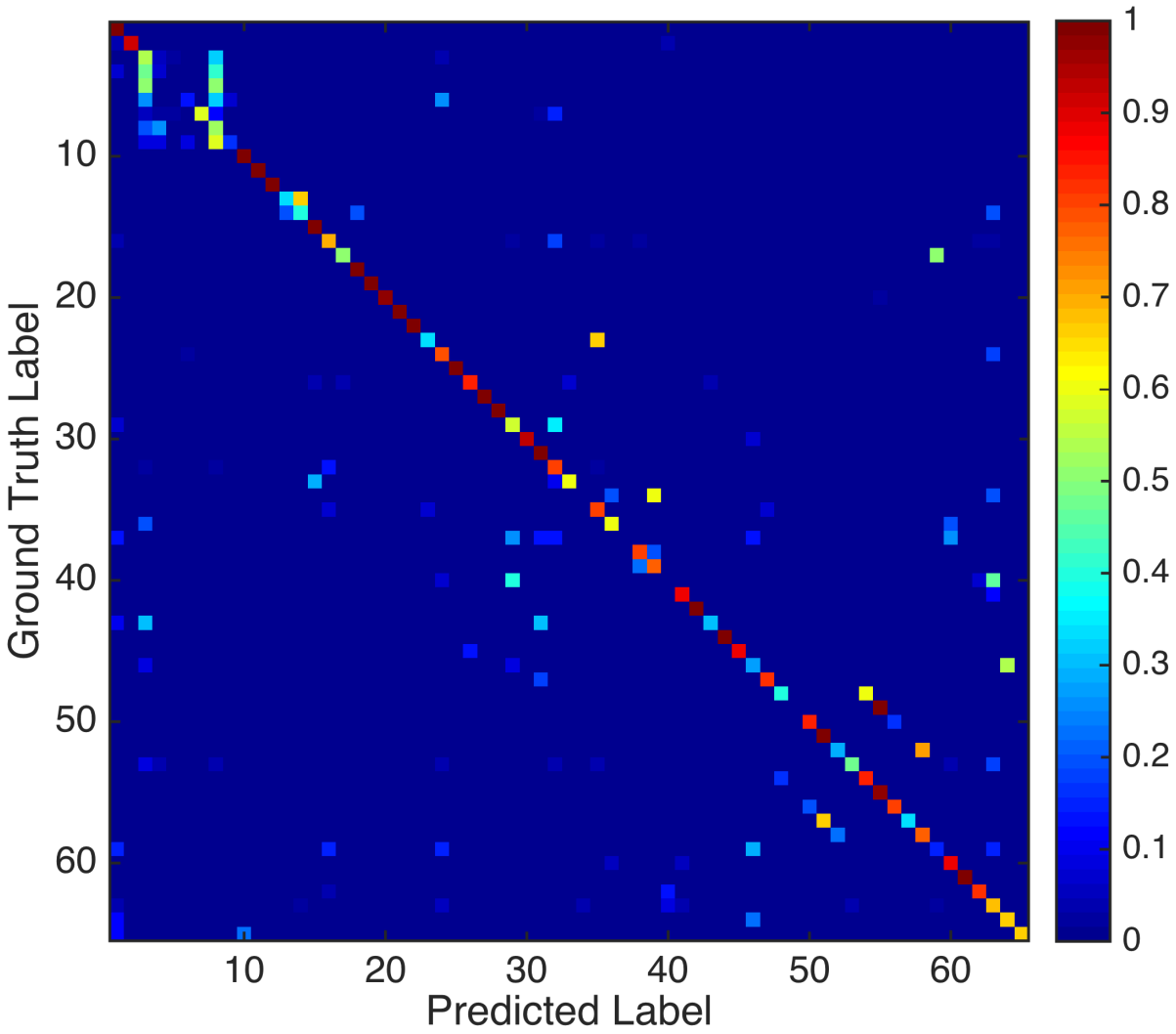

Figure 1. Confusion matrix for label prediction. Only a few off-diagonal elements have higher values (red) which indicates that in most of the cases the predicted label matched the ground truth label (higher values of diagonal elements). These few off-diagonal high values correspond to similar activity pairs, for example one off-diagonal high value is seen in row 49, column 55 indicating the predicted label to be 'take out from drawer' where the corresponding diagonal element (row 49, column 49) corresponds to the label 'take and put in drawer'.

Go to Table of Contents

2. More Example Activity Sequences Similar to Figure 4 of the Paper

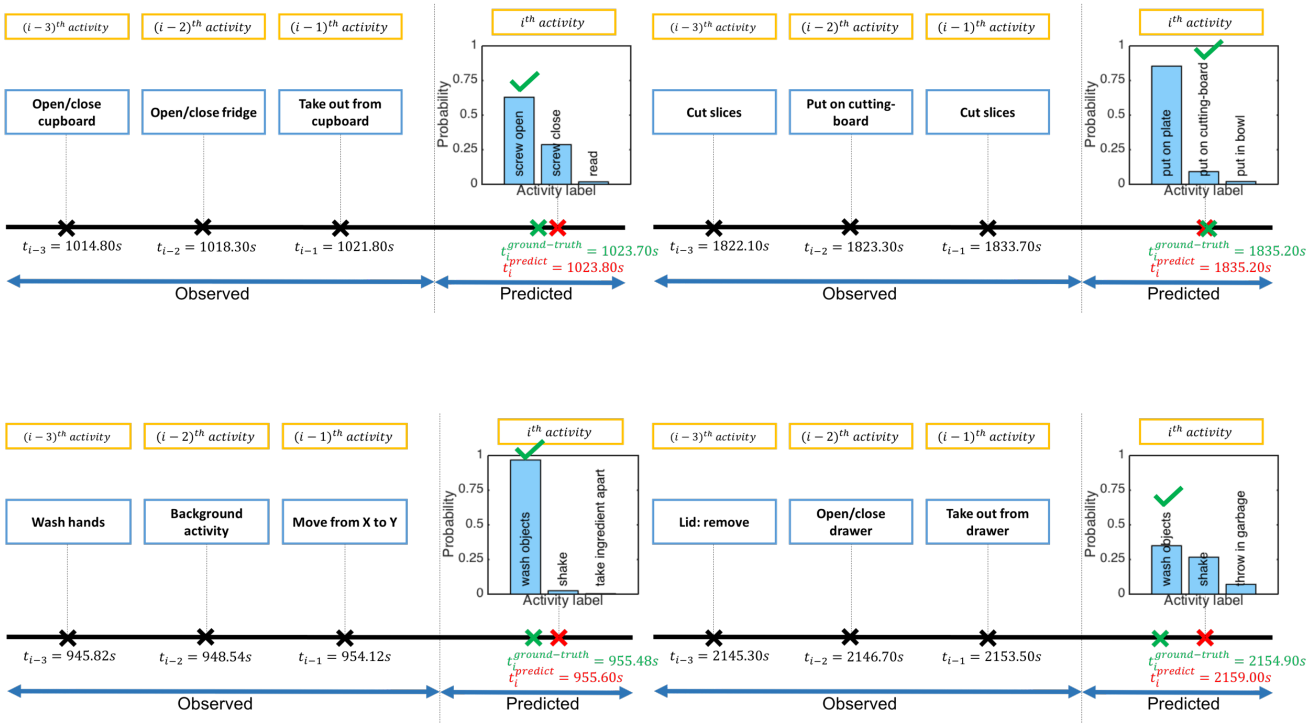

Figure 2. Four example activity sequences showing our label prediction results and time prediction results. For time prediction, green $\times$ marks the ground truth starting time of the activity we are trying to predict, and red $\times$ marks the predicted time. For label prediction, top-3 matches are shown here and in most of the cases our top-1 match corresponds to the activity that actually happened (green tick).

Go to Table of Contents

3. Future Activity Starting Time Prediction

3.1. MPII-Cooking Dataset: Activity Class-wise RMSE

Activity Class	RMSE (sec)	Activity Class	RMSE (sec)
1	0.78	34	0.68
2	0.25	35	0.89
3	0.89	36	0.19
4	1.06	37	1.07
5	1.06	38	0.45
6	0.60	39	0.36
7	0.89	40	0.47
8	1.60	41	0.27
9	0.67	42	1.48
10	0.53	43	2.07
11	1.30	44	3.31
12	3.46	45	3.52
13	0.08	46	1.55
14	0.67	47	0.45
15	0.26	48	0.47
16	0.46	49	0.35
17	0.93	50	0.51
18	3.85	51	1.11
19	0.74	52	0.20
20	0.51	53	0.89
21	1.45	54	2.28
22	0.23	55	0.86
23	0.77	56	1.31
24	3.23	57	0.68
25	0.64	58	0.26
26	0.90	59	0.28
27	0.79	60	0.61
28	9.63	61	3.18
29	0.93	62	0.59
30	0.76	63	0.74
31	0.96	64	1.87
32	0.44	65	0.86
33	0.68		

Table 2. RMSE values based on the label of the last observed activity. This table corresponds to Figure 5 (top) of the paper.

Go to Table of Contents

3.2. VIRAT Ground Dataset: Activity Class-wise RMSE

Activity Class	RMSE (sec)
1	8.01
2	6.60
3	10.43
4	6.18

Table 3. RMSE values based on the label of the last observed activity. The reason we consider these four activity classes is mentioned in the paper.

4. Multi-step Prediction

4.1. Multi-step Joint Prediction

Dataset	1-Step Prediction Accuracy (%)	2-Step Prediction Accuracy (%)	3-Step Prediction Accuracy (%)
MPII-Cooking Dataset	80.1	78.2	77.6
VIRAT Ground Dataset	71.8	70.7	68.6

Table 4. Accuracy of the predicted labels for multi-step prediction. This table corresponds to Figure 6 (top) of the paper. For both of the datasets, the label prediction accuracy decreases as we try to predict further ahead as expected.

Dataset	1-Step Prediction RMSE (sec)	2-Step Prediction RMSE (sec)	3-Step Prediction RMSE (sec)
MPII-Cooking Dataset	1.25	10.46	16.96
VIRAT Ground Dataset	10.46	-	-

Table 5. RMSE of the predicted starting times for multi-step prediction. This table corresponds to Figure 6 (bottom) of the paper. The reason we did not perform multi-step time prediction for VIRAT Ground Dataset is mentioned in the paper. The RMSE for predicted times increases with the increasing forecasting horizon as expected.

Go to Table of Contents

5. Dataset Details

5.1. MPII-Cooking Dataset

Description	This dataset contains fine-grained complex cooking activities in an indoor setting. 65 different cooking activities were performed by 12 participants ranging from beginner cooks to experienced chefs. They prepared one to six of a total of 14 dishes (sandwich, salad, fried potatoes, potato pancake, omelet, soup, pizza, casserole, mashed potato, snack plate, cake, fruit salad, cold drink, and hot drink). The activities have low inter-class variability and high intra-class variability because of diverse subjects and cooking ingredients. The amount of occlusion, clutter, or viewpoint change is small.
Number of Subjects	12
Number of Videos	44
Number of Activities	5609
Number of Activity Classes	65
Activity Types	background activity, change temperature, cut apart, cut dice, cut in, cut off ends, cut out inside, cut slices, cut stripes, dry, fill water from tap, grate, lid: put on, lid: remove, mix, move from X to Y, open egg, open tin, open/close cupboard, open/close drawer, open/close fridge, open/close oven, package X, peel, plug in/out, pour, pull out, puree, put in bowl, put in pan/pot, put on bread dough, put on cutting board, put on plate, read, remove from package, rip open, scratch off, screw close, screw open, shake, smell, spice, spread, squeeze, stamp, stir, strew, take & put in cupboard, take & put in drawer, take & put in fridge, take & put in oven, take & put in spice holder, take ingredient apart, take out from cupboard, take out from drawer, take out from fridge, take out from oven, take out from spice holder, taste, throw in garbage, unroll dough, wash hands, wash objects, whisk, wipe clean.
Associated Objects	Bowl, orange, hand, bottle, can opener, drawer, fridge, knife, etc.
Video Resolution	1624×1224
Total Video Duration	8 hours 20 minutes
Is it Wild?	No
Is it Segmented?	No
How is the Background?	Fixed

Figure 3. Some example frames from MPII-Cooking Dataset.

Go to Table of Contents

5.2. VIRAT Ground Dataset

Description	This is a state-of-the-art human activity dataset with wide variation in the activities and a high amount of clutter and occlusion. It consists of surveillance videos such as parking lot videos involving interactions between persons, vehicles and objects. The scenes are captured on a single camera but the viewpoint can differ from one scene to the next. Activities occur at different orientations based on the location and persons of interest are usually far away from the camera as these are wide-area videos.
Number of Scenes	11
Number of Videos	329
Number of Activities	1555
Number of Activity Classes	11
Activity Types	Person loading an object, person unloading an object, person opening a vehicle trunk, person closing a vehicle trunk, person getting into a vehicle, person getting out of a vehicle, person gesturing, person carrying an object, person running, person entering a facility, and person exiting a facility.
Associated Objects	Person, Vehicle, etc.
Video Resolution	1920×1080
Total Video Duration	About 5 hours
Is it Wild?	Yes
Is it Segmented?	No
How is the Background?	Fixed for a sequence

Figure 4. Some example frames from VIRAT Ground Dataset.

Go to Table of Contents

References

[1] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1194–1201. 1

[2] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3153–3160. 1