Recognizing Manipulation Actions from State-Transformations

Nachwa ABOUBAKR, James L. CROWLEY, Rémi RONFARD
Univ. Grenoble Alpes, INRIA, Grenoble INP, CNRS, Laboratoire LIG, LJK

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

State Transformations

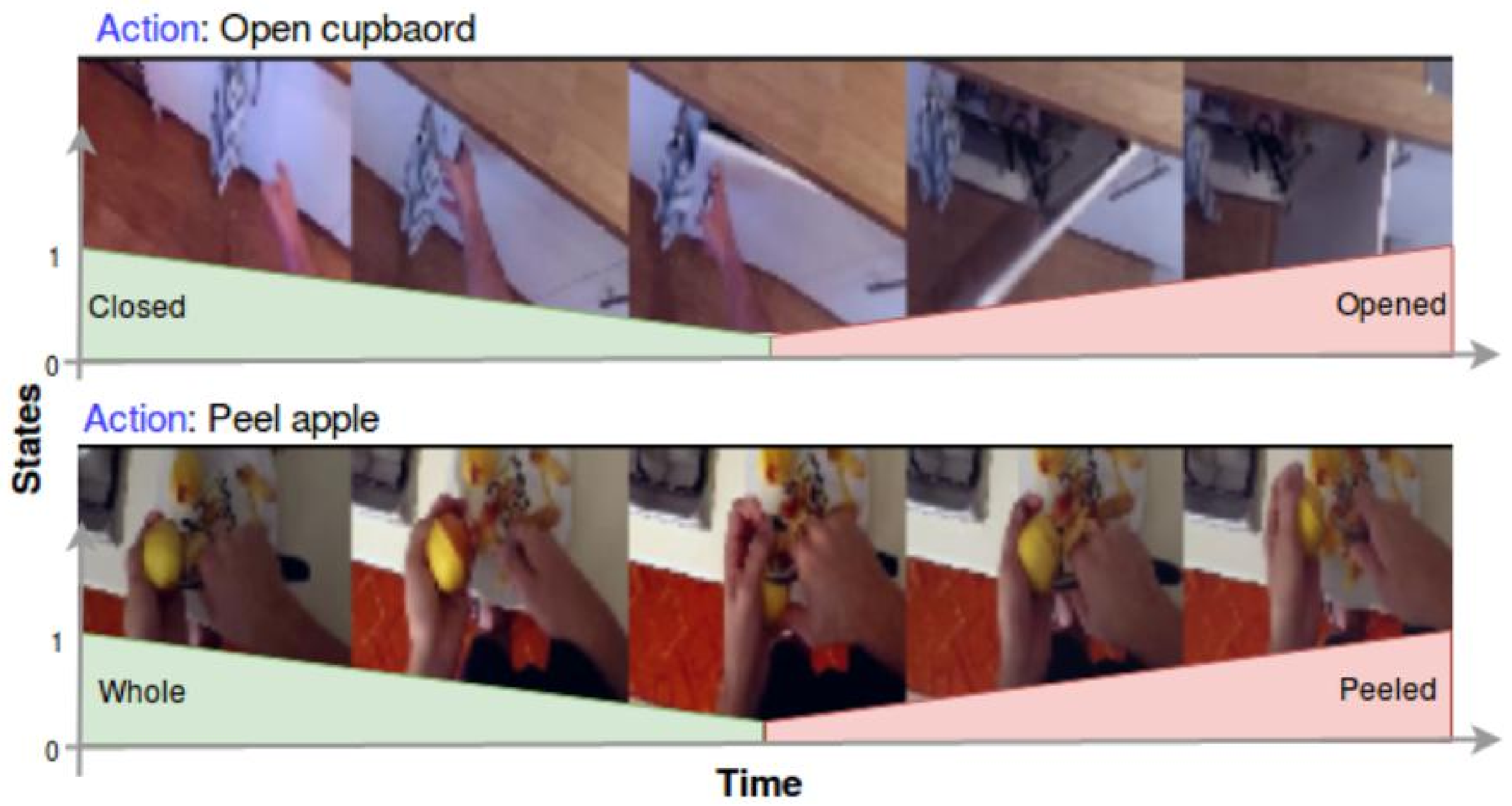


Fig1: Changes in object states over time for action recognition. Two sample sequences from the EPIC kitchen dataset.

State-Changing Actions

- State of objects are *more apparent* from still images than verbs.
- Actions can change:
  - Object's appearance,
  - Object's shape,
  - Object's position.

State Transition

$V_i : S_{before} \rightarrow S_{after};$ where $V_i \in$ verbs, $S_j \in$ states.

Our Model:

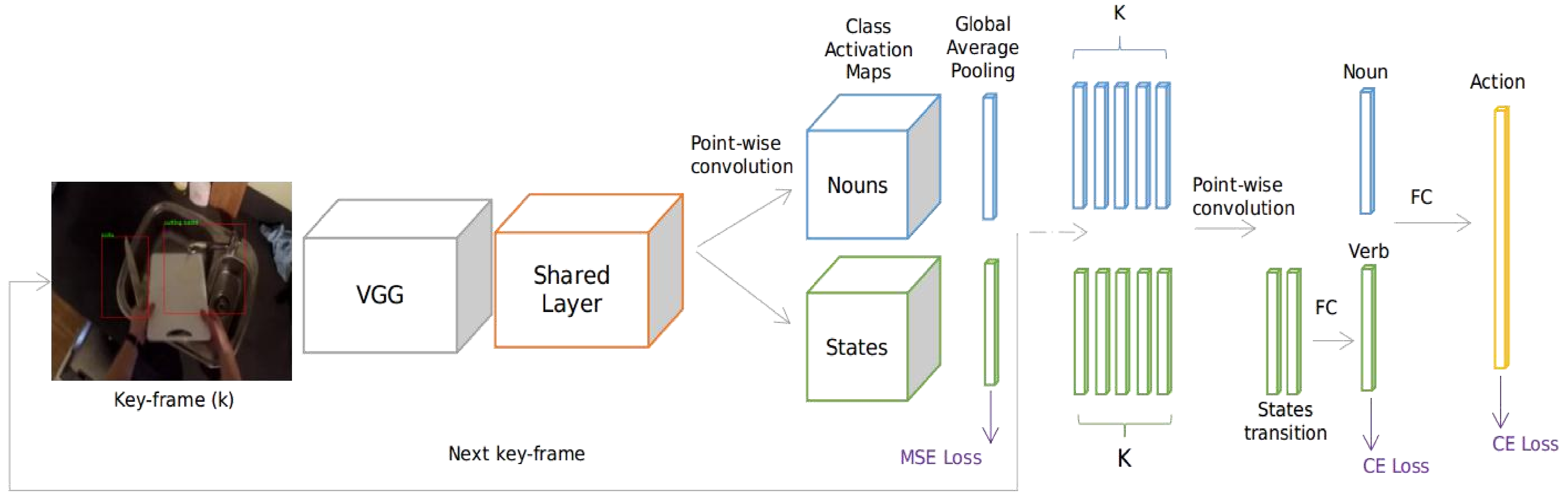


Fig 2: Proposed model to learning action recognition as state transformation.

Experiement on EPIC-Kitchens Dataset

- We defined 49 state transitions from 31 states.

a. Action Recognition Challenge.

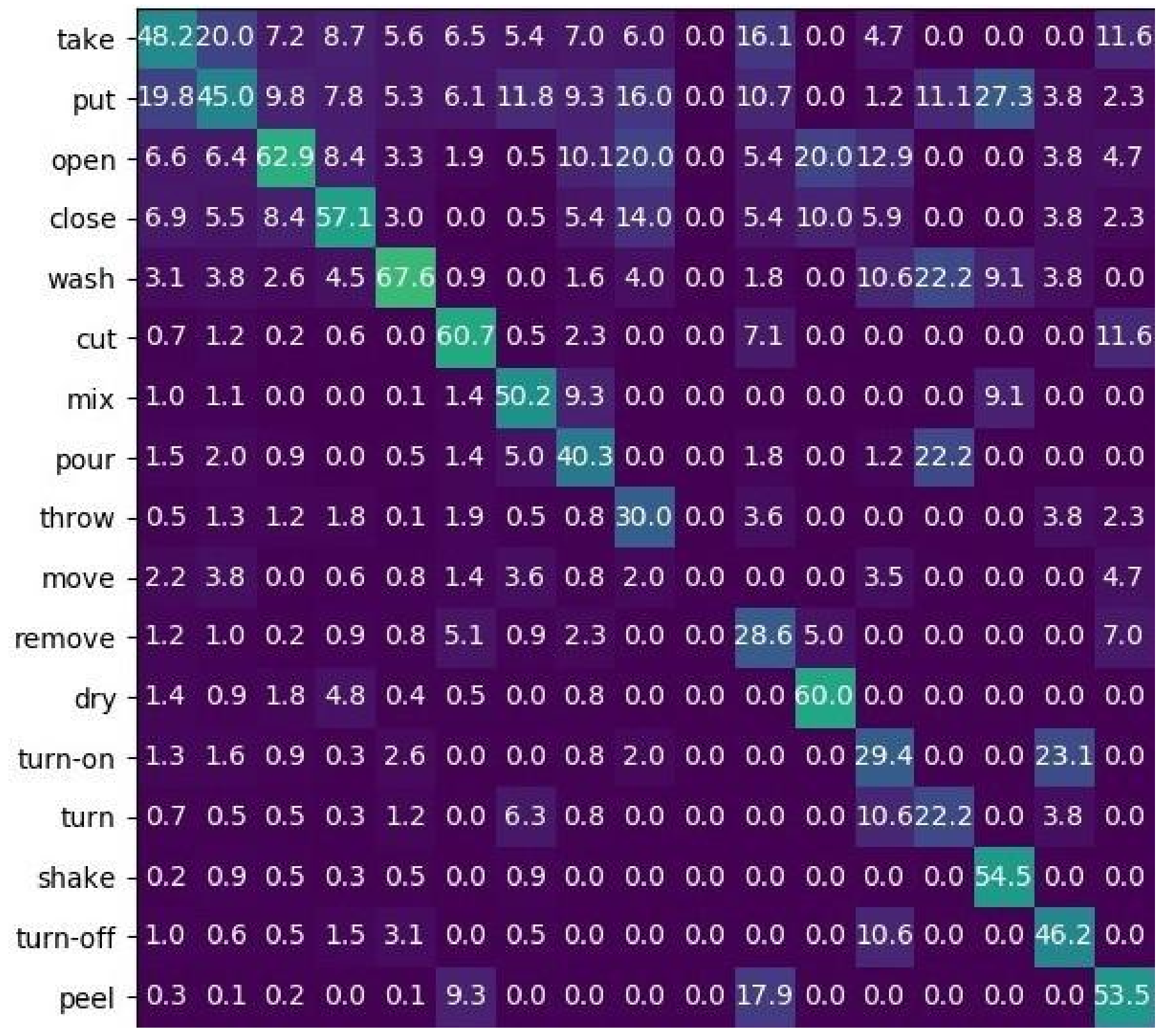
Verbs Results				
	Seen kitchens subset (**S1**)			
	Acc T1	Acc T5	Precision	Recall
Our model(RGB)	47.41	81.33	31.20	20.43
2SCNN[2](RGB)	40.44	83.04	33.74	15.9
TSN[3](RGB)	45.68	85.56	61.64	23.81
	Unseen kitchens subset (**S2**)			
Our model(RGB)	34.35	69.24	15.09	11.00
2SCNN[2](RGB)	33.12	73.23	16.06	9.44
TSN[3](RGB)	34.89	74.56	19.48	11.22

Table 1: Comparison of our method and baseline methods reported by [1].

b. Results on state-changing verbs on validation set.

	take	put	open	close	wash	cut	mix	pour	peel	Avg
Precision (%)	56.7	59.3	58.8	39.8	80.1	74.7	68.9	39.1	37.7	57.23
Recall (%)	48.2	45.0	62.9	57.1	67.7	60.7	50.2	40.3	53.5	53.96

Table 2: Our model performance on validation set on state-changing verbs.



take	48.2	20.0	7.2	8.7	5.6	6.5	5.4	7.0	6.0	0.0	16.1	0.0	4.7	0.0	0.0	0.0	11.6
put	19.8	45.0	9.8	7.8	5.3	6.1	11.8	9.3	16.0	0.0	10.7	0.0	1.2	11.1	27.3	3.8	2.3
open	6.6	6.4	62.9	8.4	3.3	1.9	0.5	10.1	20.0	0.0	5.4	20.0	12.9	0.0	0.0	3.8	4.7
close	6.9	5.5	8.4	57.1	3.0	0.0	0.5	5.4	14.0	0.0	5.4	10.0	5.9	0.0	0.0	3.8	2.3
wash	3.1	3.8	2.6	4.5	67.6	0.9	0.0	1.6	4.0	0.0	1.8	0.0	10.6	22.2	9.1	3.8	0.0
cut	0.7	1.2	0.2	0.6	0.0	60.7	0.5	2.3	0.0	0.0	7.1	0.0	0.0	0.0	0.0	0.0	11.6
mix	1.0	1.1	0.0	0.0	0.1	1.4	50.2	9.3	0.0	0.0	0.0	0.0	0.0	0.0	9.1	0.0	0.0
pour	1.5	2.0	0.9	0.0	0.5	1.4	5.0	40.3	0.0	0.0	1.8	0.0	1.2	22.2	0.0	0.0	0.0
throw	0.5	1.3	1.2	1.8	0.1	1.9	0.5	0.8	30.0	0.0	3.6	0.0	0.0	0.0	0.0	3.8	2.3
move	2.2	3.8	0.0	0.6	0.8	1.4	3.6	0.8	2.0	0.0	0.0	0.0	3.5	0.0	0.0	0.0	4.7
remove	1.2	1.0	0.2	0.9	0.8	5.1	0.9	2.3	0.0	0.0	28.6	5.0	0.0	0.0	0.0	0.0	7.0
dry	1.4	0.9	1.8	4.8	0.4	0.5	0.0	0.8	0.0	0.0	0.0	60.0	0.0	0.0	0.0	0.0	0.0
turn-on	1.3	1.6	0.9	0.3	2.6	0.0	0.0	0.8	2.0	0.0	0.0	0.0	29.4	0.0	0.0	23.1	0.0
turn	0.7	0.5	0.5	0.3	1.2	0.0	6.3	0.8	0.0	0.0	0.0	0.0	10.6	22.2	0.0	3.8	0.0
shake	0.2	0.9	0.5	0.3	0.5	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	54.5	0.0	0.0
turn-off	1.0	0.6	0.5	1.5	3.1	0.0	0.5	0.0	0.0	0.0	0.0	0.0	10.6	0.0	0.0	46.2	0.0
peel	0.3	0.1	0.2	0.0	0.1	9.3	0.0	0.0	0.0	0.0	17.9	0.0	0.0	0.0	0.0	0.0	53.5

Fig 3: Confusion Matrix on Validation set.

The code is available at:

References

[1] 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.

[2] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems, pages 568–576, 2014.

[3] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In European conference on computer vision, pages 20–36. Springer, 2016.