

Goal and overview

The problem: Relate manipulation actions and object states and discover them automatically from videos.

- Input:**
- A set of **clips** containing the same **action**.
 - An **object detector** for the class of interest.

- Output:**
- Precise temporal localization of the action.
 - Spatial and temporal localization of states.

State 1 → Action → State 2



Challenges:

- No temporal labels for object states and actions.
- Variability in appearance and motion.

Contributions:

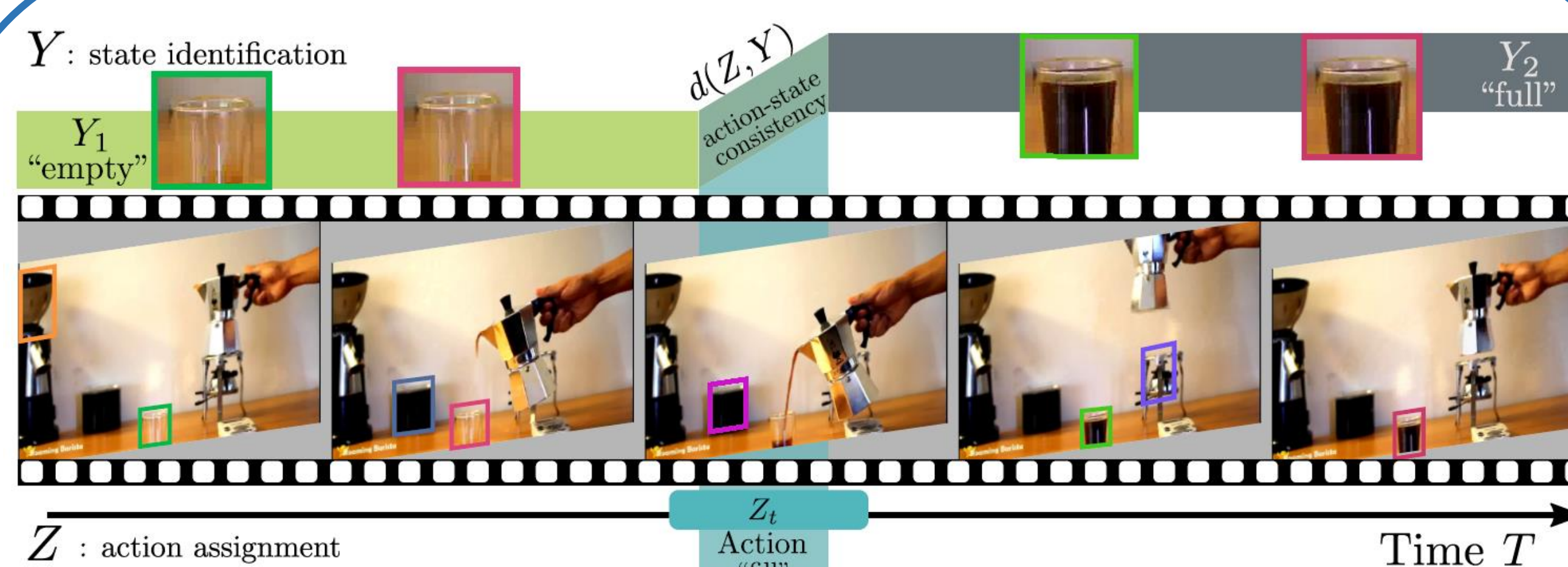
- A joint model for object states discovery and actions localization.
- An effective non-convex optimization algorithm for learning the model.
- Promising results on a challenging dataset of instructional videos.

Dataset action/states

- 7 actions, ~20-30s per video,
- Time annotation for actions,
- Track level annotation for states with labels: **state 1 | state 2 | false positive | ambiguous**
- Video extracted from YouTube, Instruction videos [2] and Charades [3].

Objects	Actions (#clips)	States	#Tracklets
wheel	{ remove (47), put (46)}	{ <i>attached, detached</i> }	5447
coffee cup	{ fill (57)}	{ <i>full, empty</i> }	1819
flower pot	{ put plant (27)}	{ <i>full, empty</i> }	2463
fridge	{ open (234), close (191)}	{ <i>open, closed</i> }	7968
oyster	{ open (28)}	{ <i>open, closed</i> }	1802

Approach



Model

$$\begin{aligned} & \text{minimize} && f(Z) + g(Y) + d(Z, Y) \\ & \text{s.t.} && Z \in \mathcal{Z} \quad \text{and} \quad Y \in \mathcal{Y} \\ & && \text{saliency of action} \quad \text{ordering + non overlap} \\ & && \text{Action localization} \quad \text{Object state labeling} \end{aligned}$$

Action cost function [1]: $f(Z) = \min_{W_v \in \mathbb{R}^{d_v \times 2}} \frac{1}{2T} \|Z - X_v W_v\|_F^2 + \frac{\lambda}{2} \|W_v\|_F^2$

Discovers temporal localization [Tx1] matrix. Representation of video [IDTF, CNN] [Tx1] matrix. Linear action classifier [dx1] matrix. Discriminative loss on data. Regularizer.

Action constraint \mathcal{Z} : One time interval is selected (saliency of action).

State cost function [1]: $g(Y) = \min_{W_s \in \mathbb{R}^{d_s \times 2}} \frac{1}{2M} \|Y - X_s W_s\|_F^2 + \frac{\mu}{2} \|W_s\|_F^2$

Discovers states [Mx2] matrix. Representation of states (CNN) [Mxd] matrix. Linear state classifier [dx2] matrix. Discriminative loss on data. Regularizer.

State constraints \mathcal{Y} :

- “Non overlap”: only one object manipulated at a time,
- Ordering constraints: State 1 → State 2,
- At least one constraint.

Joint cost: Action should be in between initial and final state.

$$d(Z_n, Y_n) = \sum_{y \in S_1(Y_n)} [t_{y_j} - t_{Z_n}]_+ + \sum_{y \in S_2(Y_n)} [t_{Z_n} - t_{y_j}]_+$$

Time of track in state 1. Time of the action. Time of track in state 2.

Optimization

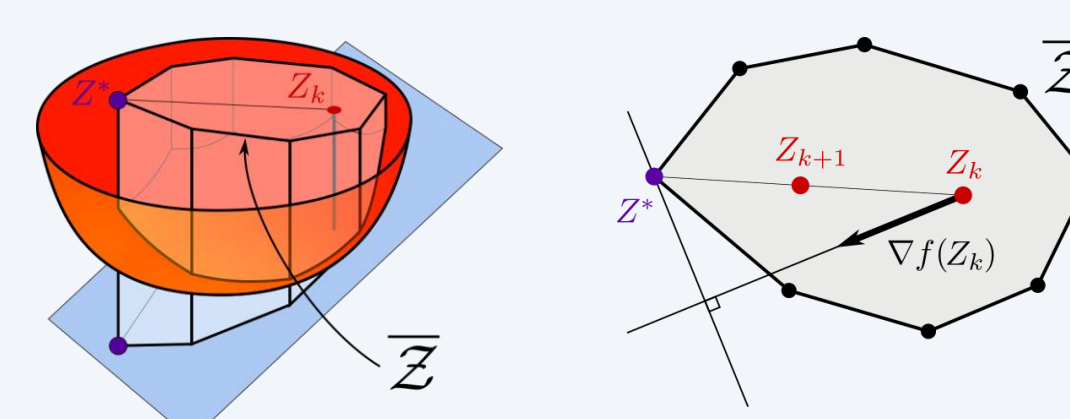
Relaxation: minimize $f(Z) + g(Y) + d(Z, Y)$ s.t. $Z \in \tilde{\mathcal{Z}}$ and $Y \in \tilde{\mathcal{Y}}$

Joint cost bilinear relaxation:

$$d(Z_n, Y_n) = \sum_{i=1}^{M_n} \sum_{t=1}^{T_n} ((Y_n)_{i1} Z_{nt} [t_{ni} - t]_+ + (Y_n)_{i2} Z_{nt} [t - t_{ni}]_+)$$

⚠ Non convex objective!

- Optimization using Frank-Wolfe [4],
- Use DP as the linear oracle to handle the constraints,
- Rounding with various techniques.

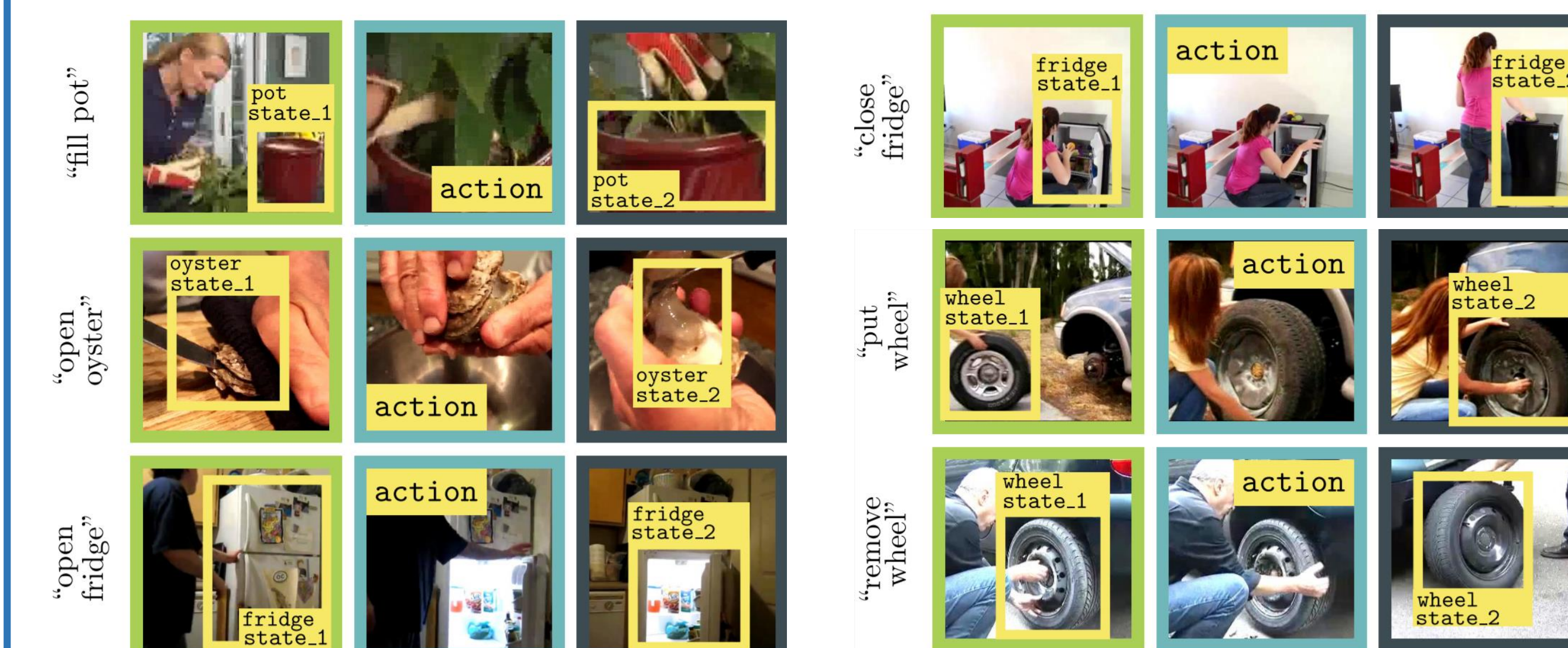


Experiments

Quantitative results

	Method	put wheel	remove wheel	fill pot	open oyster	fill coff.cup	open fridge	close fridge	Average
State discovery	(a) Chance	0.10	0.11	0.10	0.07	0.06	0.10	0.10	0.09
	(b) Kmeans	0.25	0.12	0.11	0.23	0.14	0.19	0.22	0.18
	(c) Constraints only	0.35	0.38	0.35	0.36	0.31	0.29	0.42	0.35
	(d) Salient state only	0.35	0.48	0.35	0.38	0.30	0.40	0.37	0.38
	(e) At least one state only	0.43	0.55	0.46	0.52	0.29	0.43	0.39	0.44
	(f) Joint model	0.52	0.59	0.50	0.45	0.39	0.47	0.47	0.48
	(g) Joint model + det. scores.	0.47	0.65	0.50	0.61	0.44	0.46	0.43	0.51
	(h) Joint + GT act. feat.	0.55	0.56	0.56	0.52	0.46	0.45	0.49	0.51
Action localization	(i) Chance	0.31	0.20	0.15	0.11	0.40	0.23	0.17	0.22
	(ii) [5]	0.24	0.13	0.11	0.14	0.26	0.29	0.23	0.20
	(iii) [5] + object cues	0.24	0.13	0.26	0.07	0.84	0.33	0.37	0.32
	(iv) Joint model	0.67	0.57	0.48	0.32	0.82	0.57	0.44	0.55
	(v) Joint + GT stat. feat.	0.72	0.66	0.44	0.46	0.86	0.55	0.44	0.59

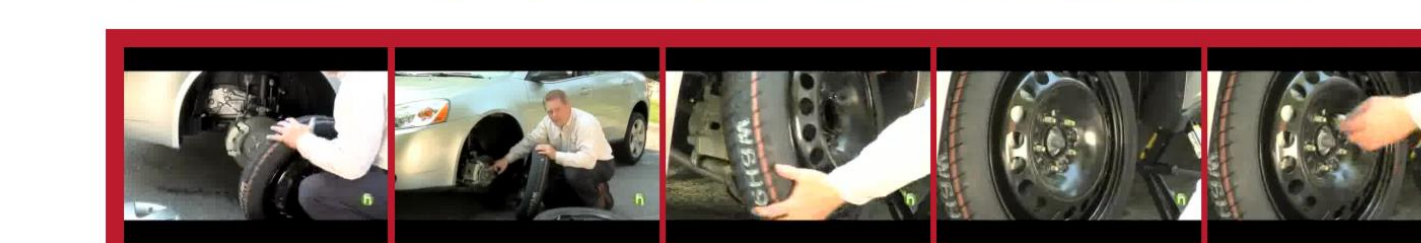
Qualitative results



Object states discovery in the wild

Obtain the clip containing manipulation action automatically from YouTube instructional videos by searching the associated narration.

Damage checking can take a while so be patient and cautious as you jack up now that the car is jacked up now we'll take the lug nuts loose remember we already pre-loosen them when the car was on the ground making it easy to notice will come right off now trying to do that while the car was jacked up would be very difficult remember to keep your lug nuts in close hand you do it by hand you do it by hand because it'll be hard to find and that's what holds your tire on again we're turning them counterclockwise to get them loose or go the left we'll remove the flat tire set it over here on the way now there was heavy traffic area you would want to leave it in the street you might want to put it to the rear of the car grab the spare tire



put it in position you're going to center the spare tire on the wheel studs that's what the lug nuts go so go ahead and do that lining it up you see it lines are pretty easy then install your lug nuts again clockwise is tight counterclockwise is loose so we'll take them up by hand as far as they'll go once the tire centered lug nuts are hand tight take the lug wrench again turning it clockwise to tighten the lugs to their firm not too tight because remember your cars up on a jack we would not want it falling off with too much leverage of force insert the jack handle back into the jack turning it counterclockwise again we're going to lower the vehicle again this will take some time but going down a lot easier than going up it seems to be going down easy you can just do it like this

	Method	put wheel	remove wheel	fill pot	open oyster	fill coff.cup	Ave.
State disc.	(e) Ctrs only	0.23	0.34	0.25	0.29	0.11	0.24
	State + det. sc.	0.33	0.48	0.28	0.40	0.13	0.32
	(g) Joint	0.38	0.53	0.25	0.43	0.20	0.36
Action local.	(g) Curated	0.63	0.68	0.63	0.63	0.53	0.62
	(i) Chance	0.14	0.10	0.06	0.10	0.15	0.11
Action local.	(ii) Action	0.05	0.10	0.00	0.15	0.25	0.11
	(iv) Joint	0.30	0.30	0.20	0.20	0.20	0.24
	(iv) Curated	0.53	0.35	0.32	0.40	0.59	0.44

References

- [1] Bach and Harchaoui. DIFFRAC: A discriminative and flexible framework for clustering. In *NIPS*, 2007.
- [2] Alayrac et al. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016.
- [3] Sigurdsson et al. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *ECCV*, 2016.
- [4] Lacoste-Julien. Convergence Rate of Frank-Wolfe for Non-Convex Objectives, Technical report, arXiv:1607.00345
- [5] Bojanowski et al. Weakly supervised action labeling in videos under ordering constraints. In *ECCV*, 2014.

Check out our project webpage for code/data!

