Joint Discovery of Object States and Manipulation Actions

Jean-Baptiste Alayrac†, Josef Sivic†*, Ivan Laptev†, Simon Lacoste-Julien‡

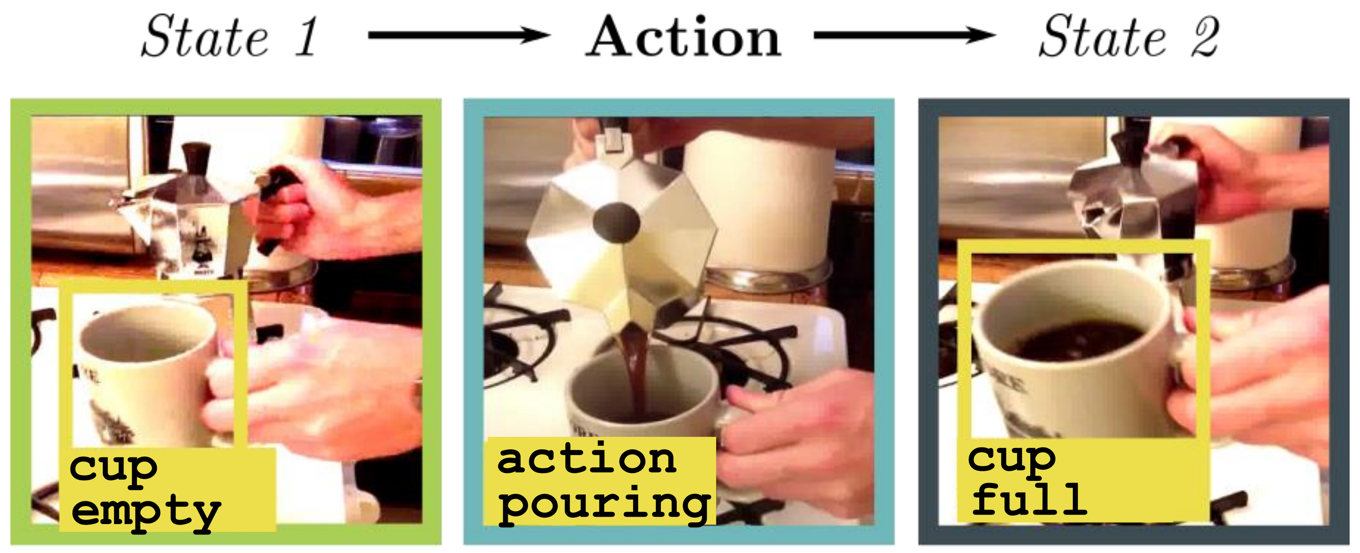
†INRIA/ENS, PSL, Paris *CIIRC, CTU in Prague ‡ MILA / DIRO, Université de Montréal

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.



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

Contributions:

i. A joint model for object states discovery and actions localization.

ii. An effective non-convex optimization algorithm for learning the model.

iii. Promising results on a challenging dataset of instructional videos.

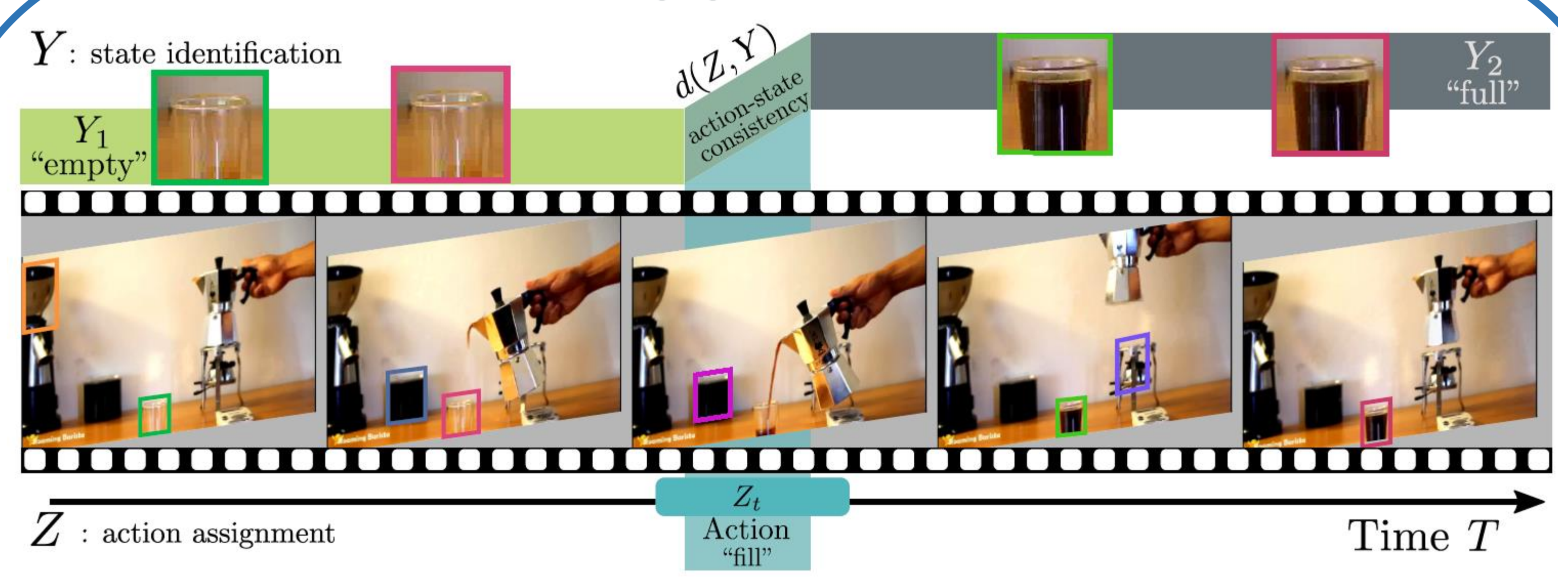
Dataset action/states

- **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)}	{*attached, detached*}	5447
coffee cup	{**fill** (57)}	{*full, empty*}	1819
flower pot	{**put plant** (27)}	{*full, empty*}	2463
fridge	{**open** (234), **close** (191)}	{*open, closed*}	7968
oyster	{**open** (28)}	{*open, closed*}	1802

Approach



Model

$$\underset{\substack{Y\in\{0,1\}^{M\times 2}\\ Z\in\{0,1\}^{T}}}{\text{minimize}} \quad \underbrace{f(Z)}_{\text{Action cost function}} + \underbrace{g(Y)}_{\text{State cost function}} + \underbrace{d(Z,Y)}_{\text{Action-state consistency}}$$

$$\text{s.t.} \quad \underbrace{Z \in \mathcal{Z}}_{\text{saliency of action}} \quad \text{and} \quad \underbrace{Y \in \mathcal{Y}}_{\text{ordering + non overlap}}$$

Action localization — **Object state labeling**

Action cost function [1]:

$$f(\underbrace{Z}_{\substack{\text{Discovered temporal localization}\\ \text{[Tx1] matrix}}}) = \min_{W_v\in\mathbb{R}^{d_v}} \underbrace{\frac{1}{2T}\|Z - \underbrace{X_v}_{\substack{\text{Representation of video (IDTF, CNN)}\\ \text{[Txd] matrix}}}\underbrace{W_v}_{\substack{\text{Linear action classifier}\\ \text{[dx1] matrix}}}\|_F^2}_{\text{Discriminative loss on data}} + \underbrace{\frac{\lambda}{2}\|W_v\|_F^2}_{\text{Regularizer}},$$

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

State cost function [1]:

$$g(\underbrace{Y}_{\substack{\text{Discovered states}\\ \text{[Mx2] matrix}}}) = \min_{W_s\in\mathbb{R}^{d_s\times 2}} \underbrace{\frac{1}{2M}\|Y - \underbrace{X_s}_{\substack{\text{Representation of states (CNN)}\\ \text{[Mxd'] matrix}}}\underbrace{W_s}_{\substack{\text{Linear state classifier}\\ \text{[dx2] matrix}}}\|_F^2}_{\text{Discriminative loss on data}} + \underbrace{\frac{\mu}{2}\|W_s\|_F^2}_{\text{Regularizer}}$$

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.

$$d(Z_n, Y_n) = \sum_{\substack{y\in\mathcal{S}_1(Y_n)\\ \text{State 1}}} [\underbrace{t_y}_{\substack{\text{Time of track in}\\ \text{state 1}}} - \underbrace{t_{Z_n}}_{\substack{\text{Time of the}\\ \text{action}}}]_+ + \sum_{\substack{y\in\mathcal{S}_2(Y_n)\\ \text{State 2}}} [t_{Z_n} - \underbrace{t_y}_{\substack{\text{Time of track in}\\ \text{state 2}}}]_+,$$

Optimization

Relaxation:

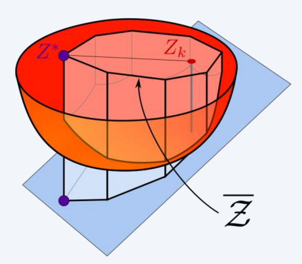
$$\underset{\substack{Y\in[0,1]^{M\times 2}\\ Z\in[0,1]^{T}}}{\text{minimize}} \quad f(Z) + g(Y) + d(Z,Y) \quad \text{s.t.} \quad Z \in \bar{\mathcal{Z}} \ \text{ and } \ Y \in \bar{\mathcal{Y}}$$

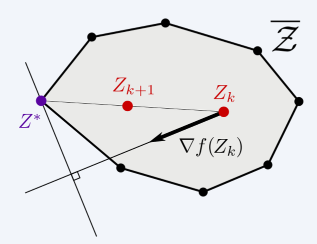
Joint cost bilinear relaxation:

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

⚠ **Non convex objective!**

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



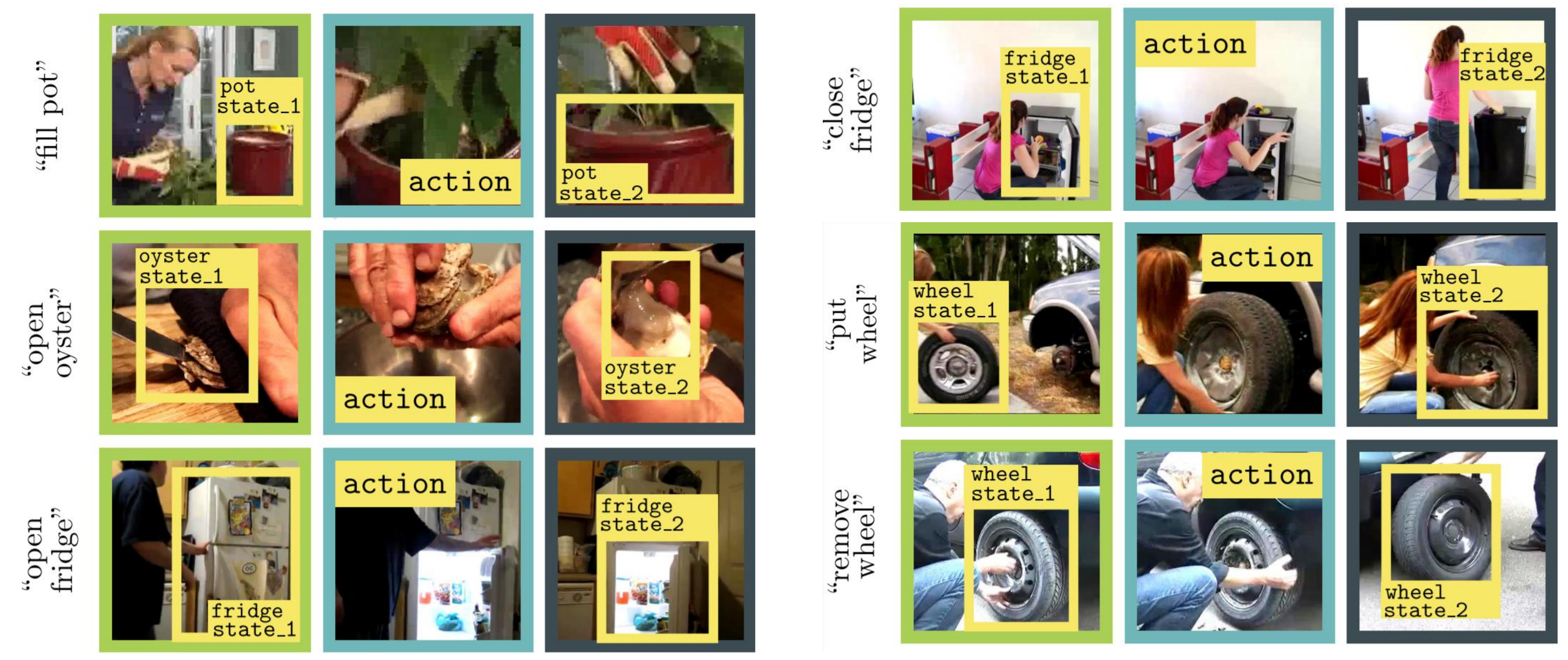


Experiments

Quantitative results

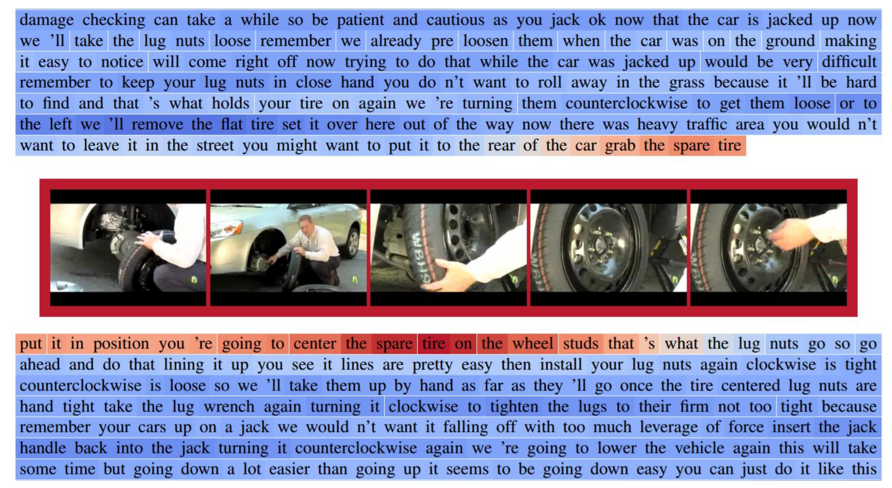
	Method	put wheel	remove wheel	fill pot	open oyster	fill coff.cup	open fridge	close fridge	**Average**
State discovery	**(a)** Chance	0.10	0.11	0.10	0.07	0.06	0.10	0.10	0.09
	(b) Kmeans	0.25	0.12	0.11	0.23	0.14	0.19	0.22	0.18
	(c) Constraints only	0.35	0.38	0.35	0.36	0.31	0.29	0.42	0.35
	(d) Salient state 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.



	Method	put wheel	remove wheel	fill pot	open oyster	fill coff.cup	**Ave.**
State disc.	**(c)** Cstrs 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**
	(g) Curated	0.63	0.68	0.63	0.63	0.53	0.62
Action local.	**(i)** Chance	0.14	0.10	0.06	0.10	0.15	0.11
	(iii) 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!