CZECH TECHNICAL
UNIVERSITY
IN PRAGUE

Learning from instructional videos

Josef Sivic

J.-B. Alayrac, N. Agrawal, I. Laptev and S. Lacoste-Julien

What are instructional videos?

[Alyarac et al., CVPR 2016]

Goal: learn a representation of a task from video

Motivation

[Darpa robot challenge]

Learning from Internet for robotics

[Microsoft HoloLens]

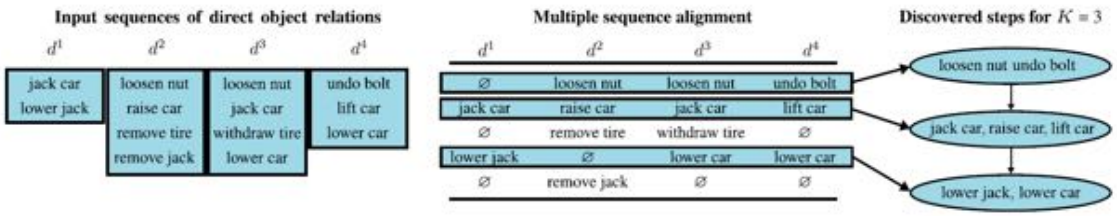
Personal assistant

Outline

1. Learn sequence of main steps of a task

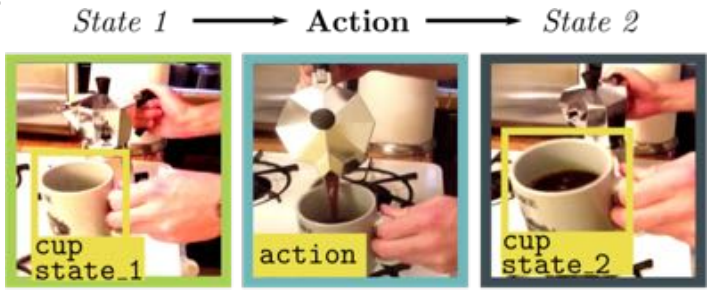
   [Alayrac et al., CVPR 2016]

   [Alayrac et al., PAMI 2017]

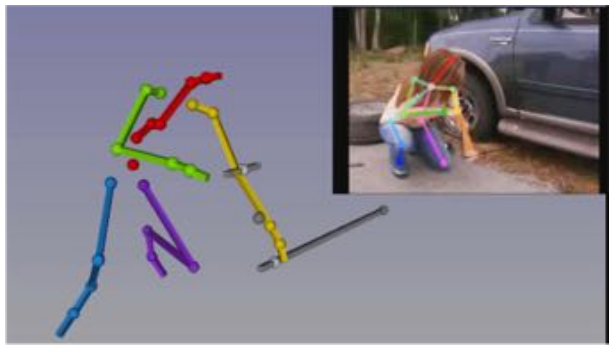


2. Modeling changes in object states

   [Alayrac et al., ICCV 2017]

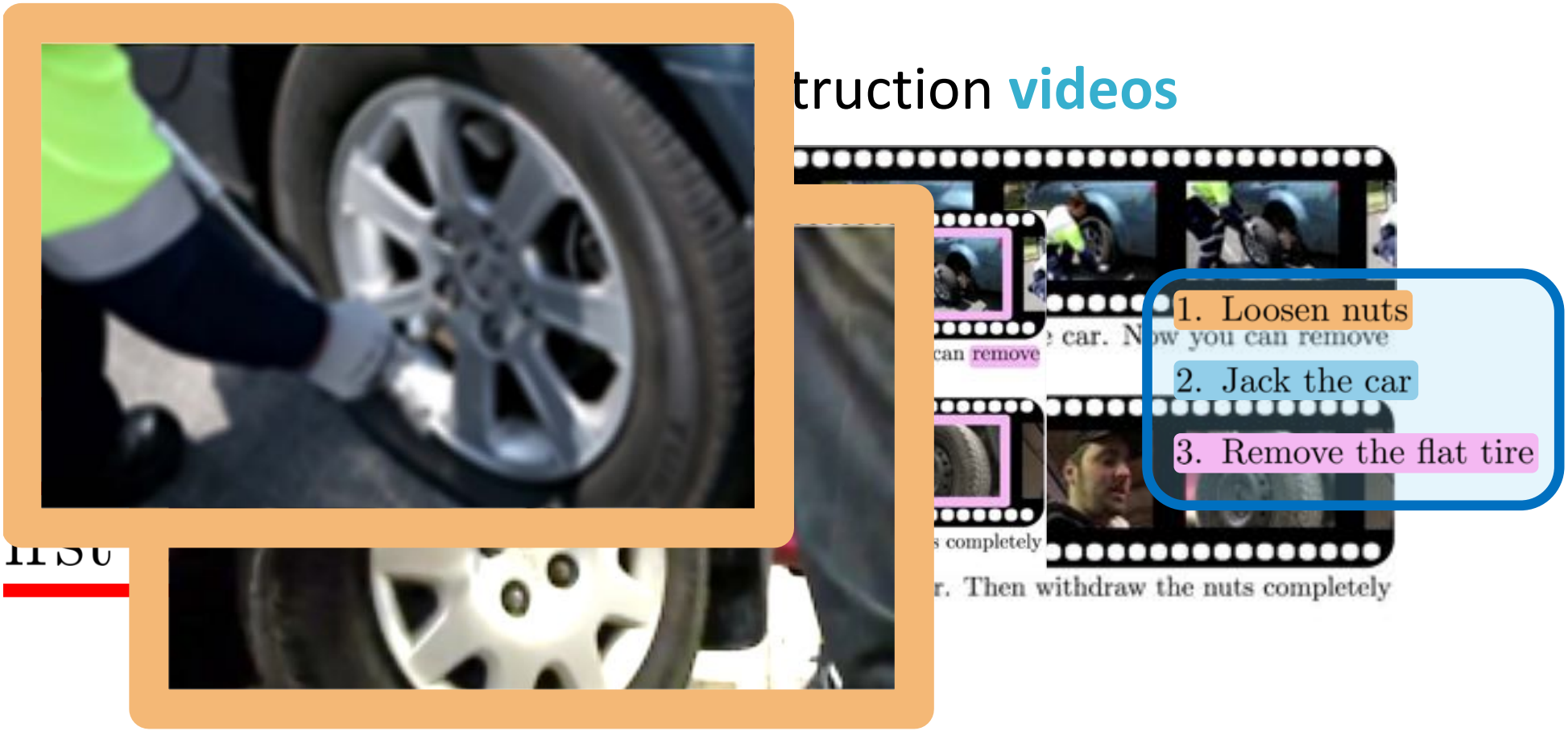


3. Discussion and challenges



Learn sequence of main steps of a task

[Alayrac et al., CVPR 2016]



Outputs:

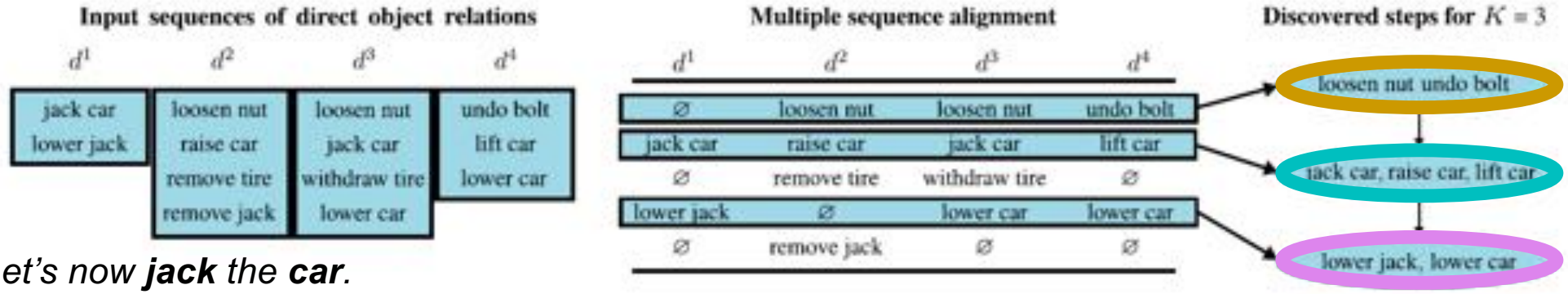
- sequence of **main steps**
- **visual** and **linguistic** representations of the steps
- temporal **localization** of each step

Related work

- **Action recognition:** [Laptev et al'08, Niebles et al'08, Schüldt'04, Wang and Schmid'13, Wang et al'11, Simonyan et al'14, Karpathy et al'14, Potapov et al'14, Donahue et al'15, Ji et al'10, Fernando et al'15, Wang et al'15, Peng et al'14, Ng et al'15, Tran et al'15, Chen et al'10, Jain et al'13, Jhuang et al'13, Khuene et al'11, Soomro et al'12, Taylor et al'10, Bregonzio et al'09, Gilbert et al'11, Ikizler-Cinbis et al'10, Kläser et al'10, Kovashka et al'10, Liu et al'09, Marszalek et al'09, Matikainen et al'09, Messing et al'09, Schüldt et al'04, Rodriguez et al'08]…
- **NLP** and **VISION**: [Chen et al'15, Duchenne et al' 09, Bojanowski et al'13,'14, Naim et al'15, Kong et al'14, Tu et al'14,Zitnick et al'13,Yan et al'15,] **image captioning** [Kulkarni et al'11, Hodosh et al'13, Vinyals et al'15, Karpathy et al '15, Lebret et al'15, Donahue et al'15, Elliot et al'14, Fang et al'15, Mao et al'14, Mitchell et al'12,], **video captioning** [Guadarrama et al'13,Rohrbach et al'13,15], **VQA** [Agrawal et al'15, Malinowski et al'15, Antol et al'15, Bigham et al'10,Gao et al'15,Ren et al'15, Yu et al'15, Zhang et al'15]…
- **Instruction videos:** [Damen et al.'14, Huang et al.'17, Kuehne et al.'15, Malmaud et al'15, Sener et al'15, Zhou et al.'17]

Approach: two linked clustering problems

1. Text clustering into a sequence of common steps



*Let's now **jack** the **car**.*

DOBJ = (jack,car)

2. Video clustering to localize the actions with text constraints

Assumptions and overview of the approach

Assumptions:

Assumption 1: Each task is composed of an **ordered** sequence of steps.

Assumption 2: People do **what** they say roughly **when** they say it

Approach:

two linked **clustering** stages

1) **Text clustering** using multiple sequence alignment

2) **Video clustering** under text constraints

Method - 1st stage:
Multiple sequence alignment of **text**

Narrations are first processed into sequence of **direct object relations** (**dobj**)

*Let's now **jack** the **car**.* → **DOBJ = (jack,car)**

A score of similarity between **dobj** is obtained from **Wordnet**

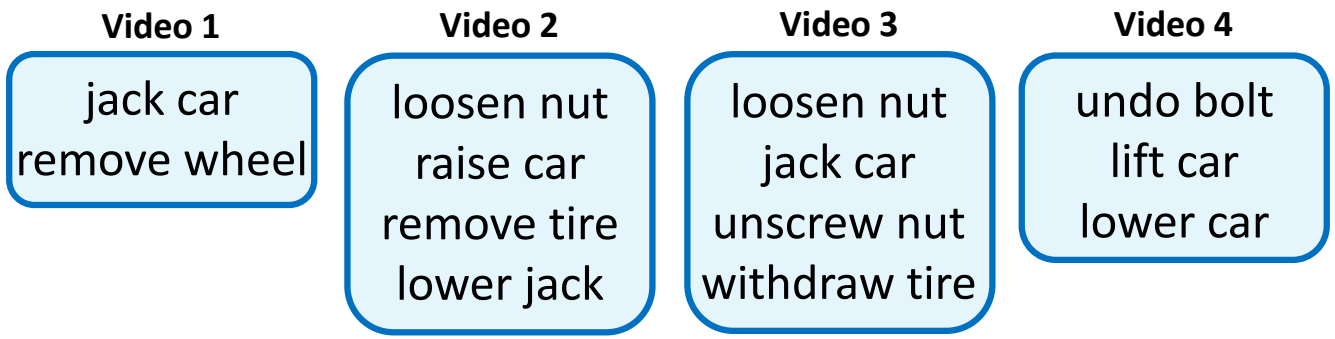
MATCH

undo bolt = loosen nut

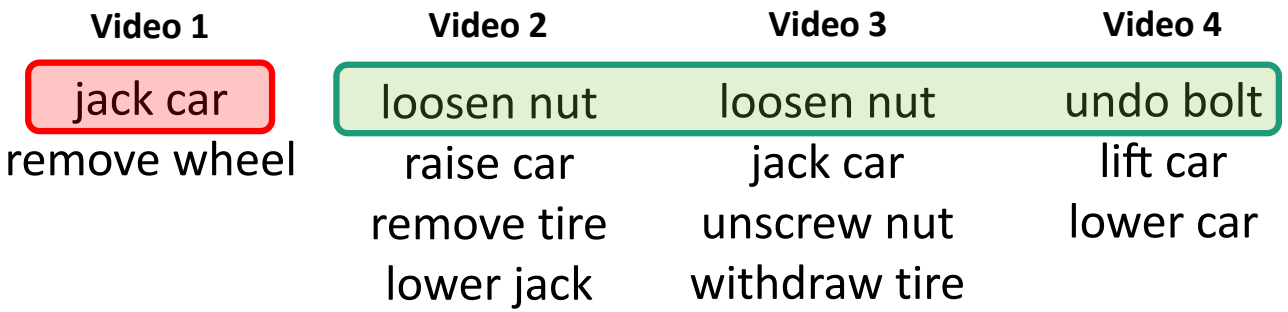
NO MATCH

jack car ≠ remove wheel

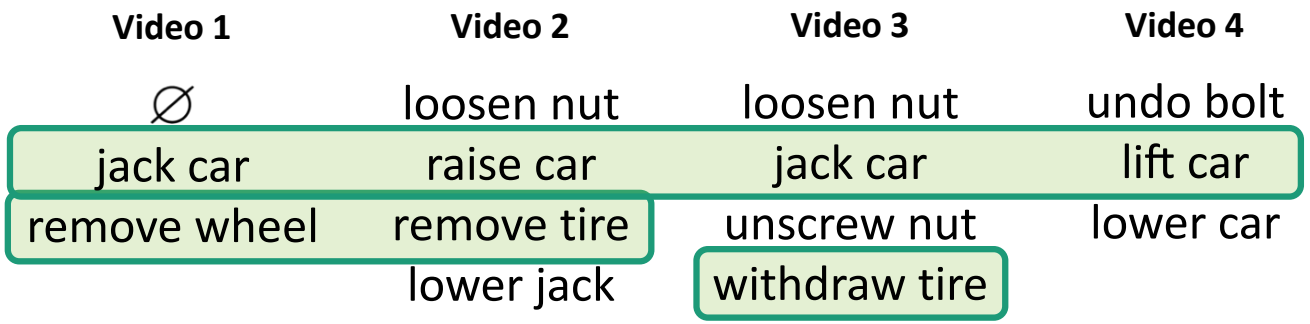
Method - 1st stage:
Multiple sequence alignment of **text**



Method- 1st stage:
Multiple sequence alignment of **text**



Method- 1st stage:
Multiple sequence alignment of **text**



Method - 1st stage:
Multiple sequence alignment of **text**

Video 1	**Video 2**	**Video 3**	**Video 4**
∅	loosen nut	loosen nut	undo bolt
jack car	raise car	jack car	lift car
remove wheel	remove tire	unscrew nut	lower car
	lower jack	withdraw tire	

Method - 1st stage:
Multiple sequence alignment of **text**

Video 1	Video 2	Video 3	Video 4
∅	loosen nut	loosen nut	undo bolt
jack car	raise car	jack car	lift car
∅	∅	unscrew nut	∅
remove wheel	remove tire	withdraw tire	lower car
∅	lower jack	∅	

Method- 1st stage:
Multiple sequence alignment of **text**

Video 1	Video 2	Video 3	Video 4
∅	loosen nut	loosen nut	undo bolt
jack car	raise car	jack car	lift car
∅	∅	unscrew nut	∅
remove wheel	remove tire	withdraw tire	∅
∅	lower jack	∅	lower car

MSA is **NP-hard**, we formulate it as a **QP** and use **Frank-Wolfe** algorithm to get an approximate solution.

[Wang and Jiang 1994, Higgins and Sharp, 1988]

Method- 1st stage:
Multiple sequence alignment of **text**

The **list of main steps** is then deduced from the alignment (here for K=3):

Video 1	Video 2	Video 3	Agreement	Video 4
∅	loosen nut	loosen nut	3	undo bolt
jack car	raise car	jack car	4	lift car
∅	∅	unscrew nut	1	∅
remove wheel	remove tire	withdraw tire	3	∅
∅	lower jack	∅	2	lower car

Method - 1st stage:
Multiple sequence alignment of **text**

The **list of main steps** is then deduced from the alignment (here for K=3):

Video 1	Video 2	Video 3	Video 4	Agreement
∅	loosen nut	loosen nut	undo bolt	**3**
jack car	raise car	jack car	lift car	**4**
∅	∅	unscrew nut	∅	~~1~~
remove wheel	remove tire	withdraw tire	∅	**3**
∅	lower jack	∅	lower car	~~2~~

Discovered list of steps

1) Loosen nut
2) Jack car
3) Remove wheel

Method - 2nd stage:
Video clustering with **text** constraints

1st step Output

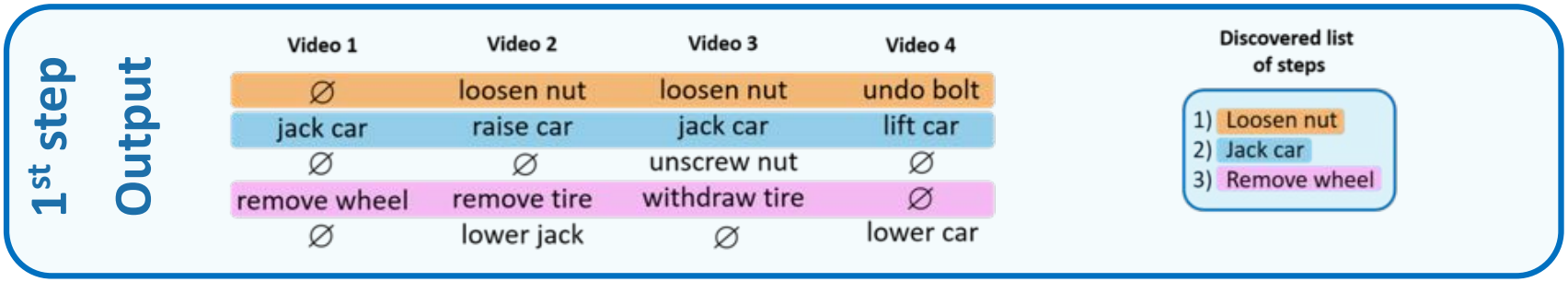
Video 1	Video 2	Video 3	Video 4
∅	loosen nut	loosen nut	undo bolt
jack car	raise car	jack car	lift car
∅	∅	unscrew nut	∅
remove wheel	remove tire	withdraw tire	∅
∅	lower jack	∅	lower car

Discovered list of steps

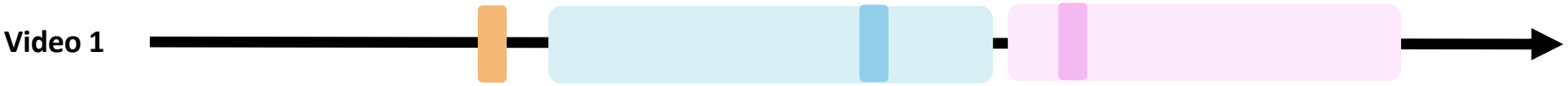
1) Loosen nut
2) Jack car
3) Remove wheel

Goal: get the **temporal localization** of the steps in each video

Method - 2nd stage:
Video clustering with **text** constraints



Goal: get the **temporal localization** of the steps in each video



Method - 2nd stage:
Video clustering with **text** constraints

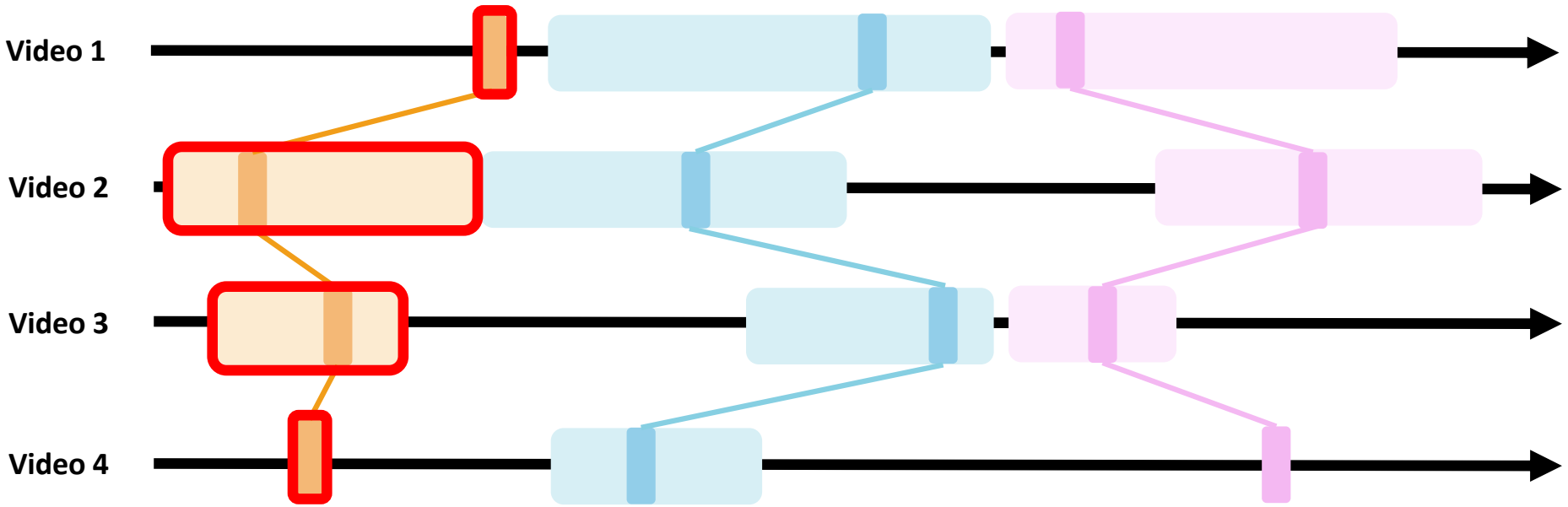
1st step Output

Video 1	Video 2	Video 3	Video 4
∅	loosen nut	loosen nut	undo bolt
jack car	raise car	jack car	lift car
∅	∅	unscrew nut	∅
remove wheel	remove tire	withdraw tire	∅
∅	lower jack	∅	lower car

Discovered list of steps

1) Loosen nut
2) Jack car
3) Remove wheel

Goal: get the **temporal localization** of the steps in each video



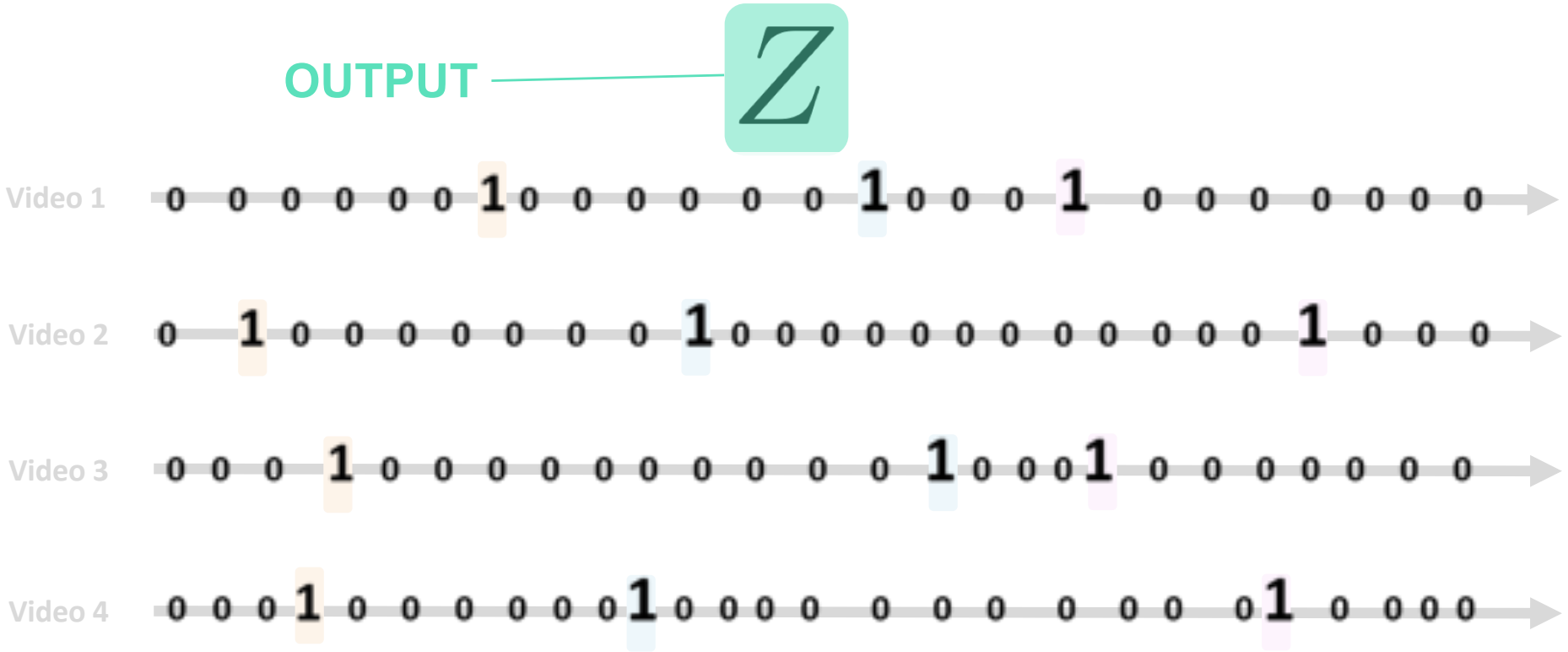
Method - 2nd stage:
Video clustering with **text** constraints

1st step Output

Video 1	Video 2	Video 3	Video 4
∅	loosen nut	loosen nut	undo bolt
jack car	raise car	jack car	lift car
∅	∅	unscrew nut	∅
remove wheel	remove tire	withdraw tire	∅
∅	lower jack	∅	lower car

Discovered list of steps

1) Loosen nut
2) Jack car
3) Remove wheel



Method - 2nd stage:
Video clustering with **text** constraints

1st step Output

Video 1	Video 2	Video 3	Video 4
∅	loosen nut	loosen nut	undo bolt
jack car	raise car	jack car	lift car
∅	∅	unscrew nut	∅
remove wheel	remove tire	withdraw tire	∅
∅	lower jack	∅	lower car

Discovered list of steps

1) Loosen nut
2) Jack car
3) Remove wheel

Method: **Discriminative clustering.**

$$h(Z) = \min_{W \in \mathbb{R}^{K \times d}} \underbrace{\frac{1}{2T}\|Z - XW\|_F^2}_{\text{Discriminative loss on data}} + \underbrace{\frac{\lambda}{2}\|W\|_F^2}_{\text{Regularizer}} .$$

OUTPUT (Discovered temporal localization)

Representation of video chunks (IDTF,CNN) [Txd] matrix

Linear action classifier [dxK] matrix

$$\min_Z \; h(Z) \quad \text{s.t.} \quad \underbrace{Z \in \mathcal{Z}}_{\text{ordered script}} , \quad \underbrace{AZ \geq R}_{\text{weak textual constraints}}$$

Subtitle alignment [SxT] matrix

Text Assignment [SxK] matrix

[Bach and Harchoui'07, Xu et al.'04, Bojanowski et al.'15]

Illustration in the feature space

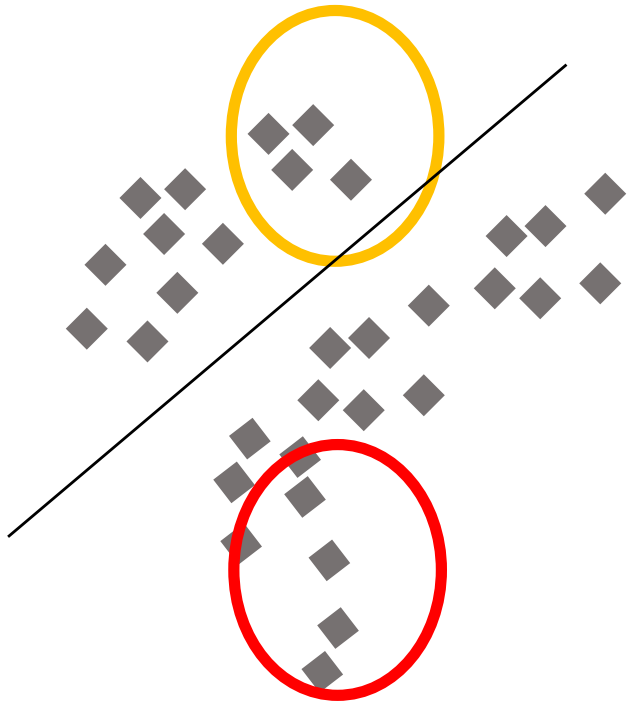
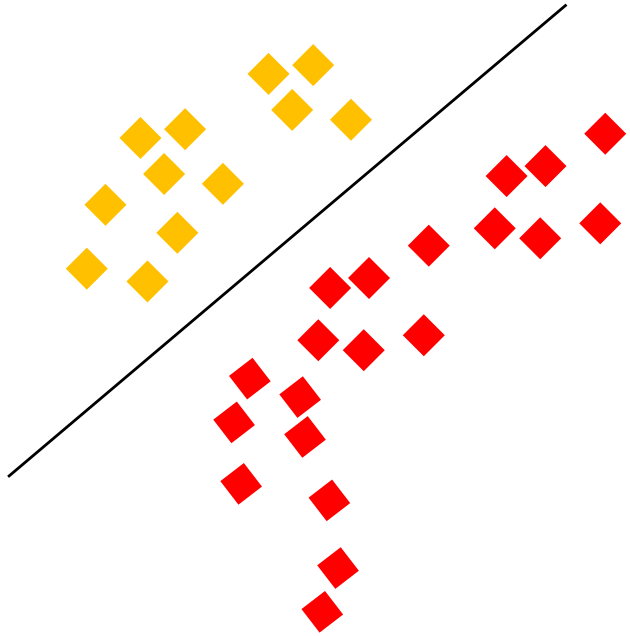


Illustration in the feature space

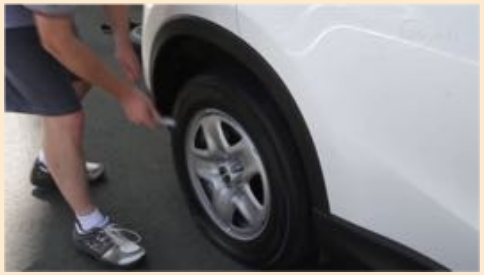


Experiments: A new dataset

- **30** videos per task
- ~**2 min** per video
- Manual correction of transcriptions from ASR
- Manual annotation of 7-10 main steps for each task **(evaluation only)**

See also [Sener et al., ICCV'15]

Qualitative results


``loosen nuts”

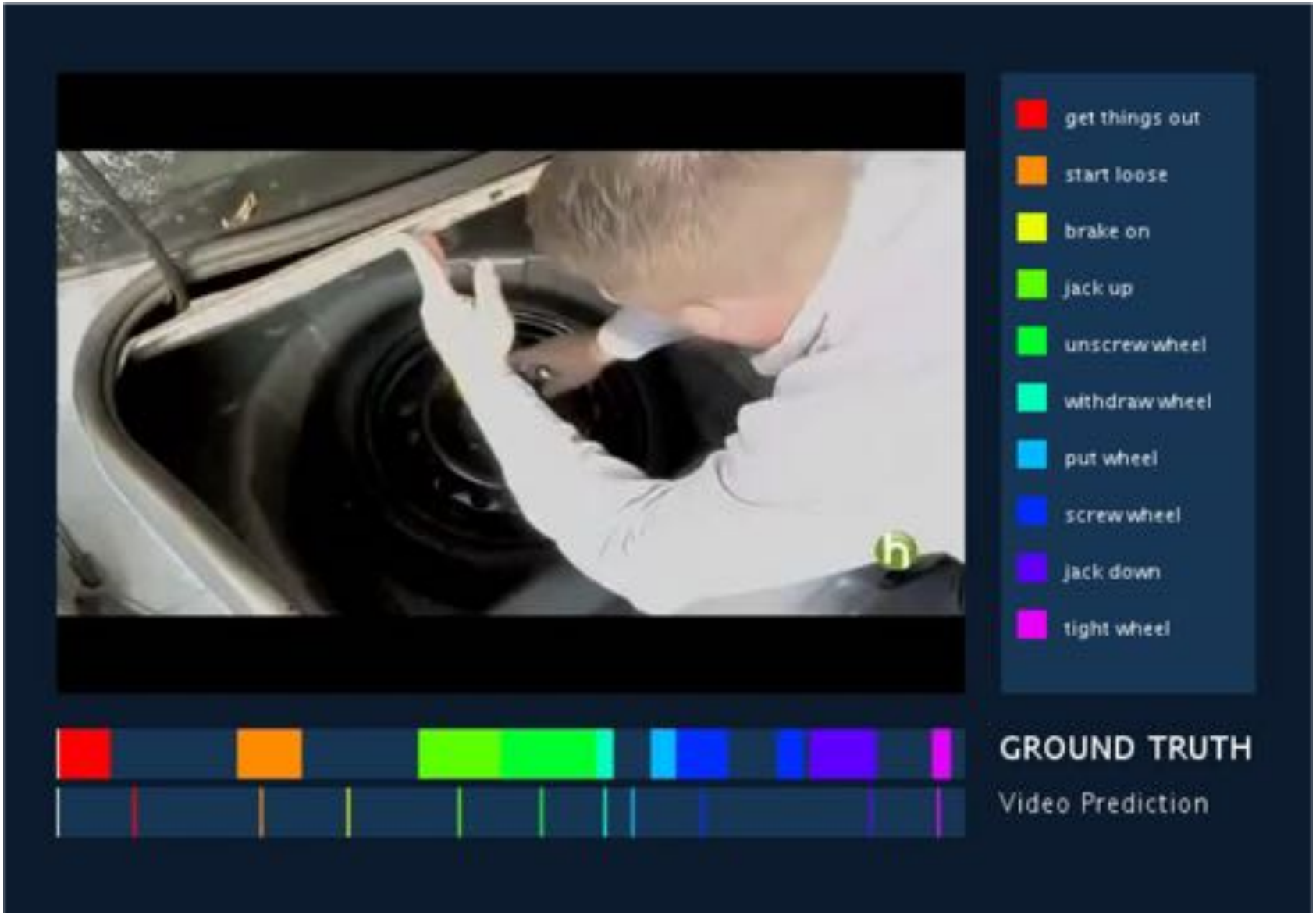


``jack car”



``remove wheel”

Qualitative results



More results

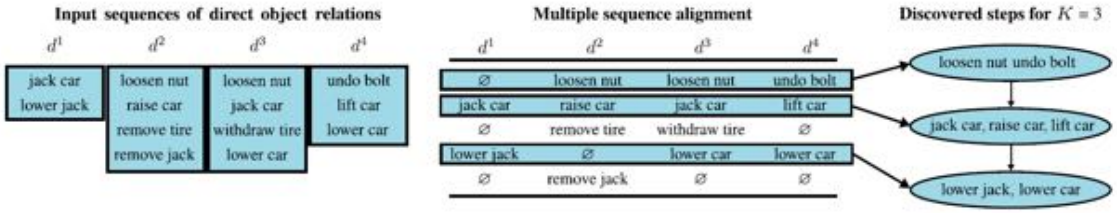
www.di.ens.fr/willow/research/instructionvideos/

Outline

1. Learn sequence of main steps of a task

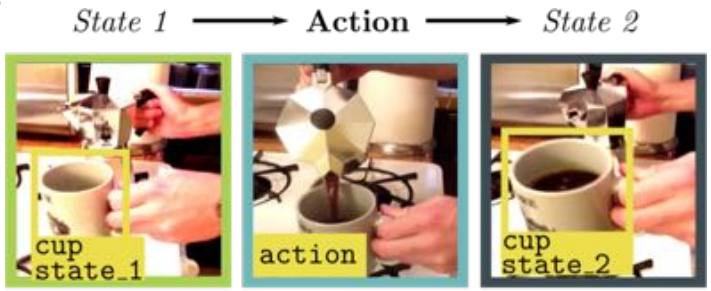
   [Alayrac et al., CVPR 2016]

   [Alayrac et al., PAMI 2017]

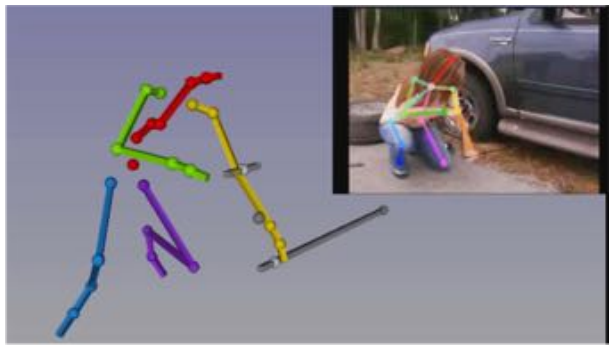


2. Modeling changes in object states

   [Alayrac et al., ICCV 2017, to appear]



3. Discussion and challenges



What about objects?

[Alayrac et al., ICCV 2017, to appear]

To complete a step, you often need to modify **state of object**.

Empty cup *State 1* → **Fill** **Action** → **Full** cup *State 2*

Also, e.g. **open** a *door,* **fill** a *water bottle,* ***cut*** *bread,…*

Can we learn the set of **actions** and **object states** from data?

The goal

Input & Output

- Set of N clips depicting the same **action**
- A pre-trained **object detector**

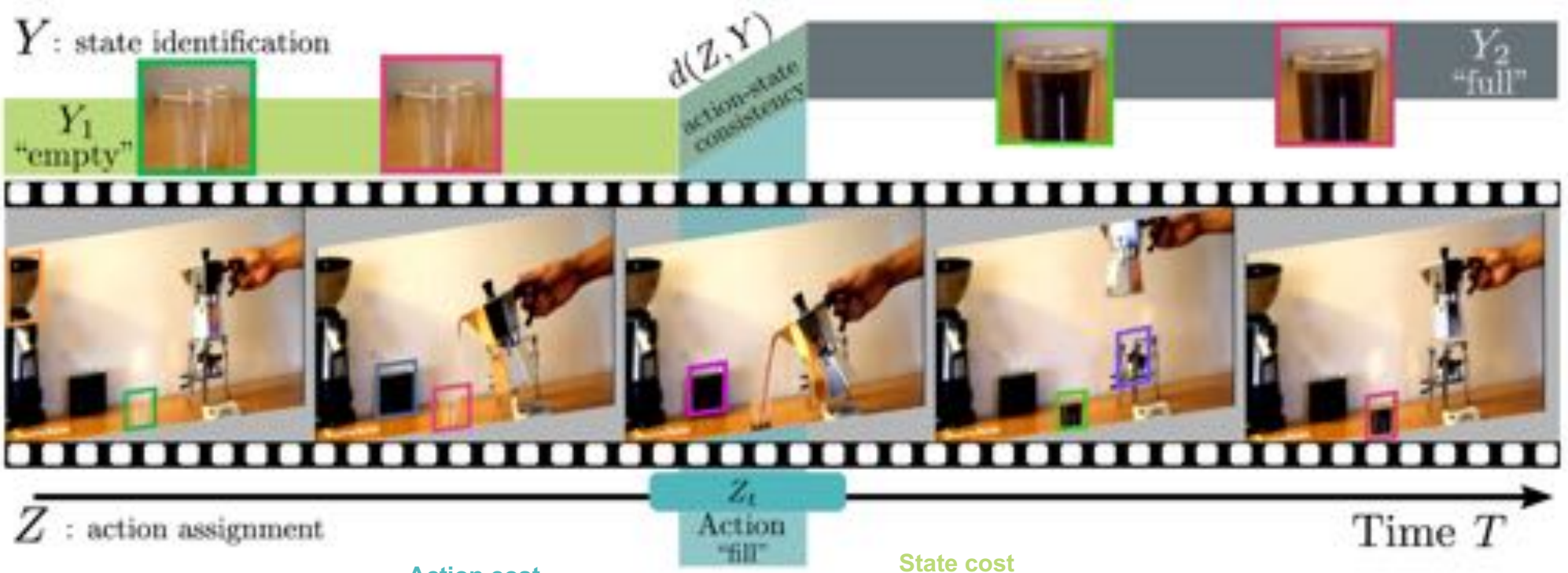
- temporal **localization** of the action
- **spatial/temporal** localization of states

Related work

- **Object attributes :** [Farhadi et al., CVPR 2009], [Parikh and Grauman, ICCV 2011], [Patterson et al., IJCV 2014], [Duan et al., CVPR 2012]
- **Object states :** [Brady, Konkle, Oliva, Alvarez 2006], [Fathi and Rehg, CVPR 2009], [Pirsiavash and Ramanan CVPR 2012], [Isola et al, CVPR 2015]
- **Action as transformations:** [Wang et al, CVPR 2016]
- **Manipulation actions and person-object interactions:** Instruction videos [Alayrac et al, CVPR 2015, Sener et al., ICCV 2015], Charades dataset [Sigurdsson, Varol et al, ECCV 2016], [Kjellstrom et al., 2011], [Gupta et al., PAMI 2009], [Gall et al., CVPR 2011], [Koppula and Saxena, ECCV 2014], …

Can we learn actions and object states with minimal supervision?

Overview of the approach



Action cost function

State cost function

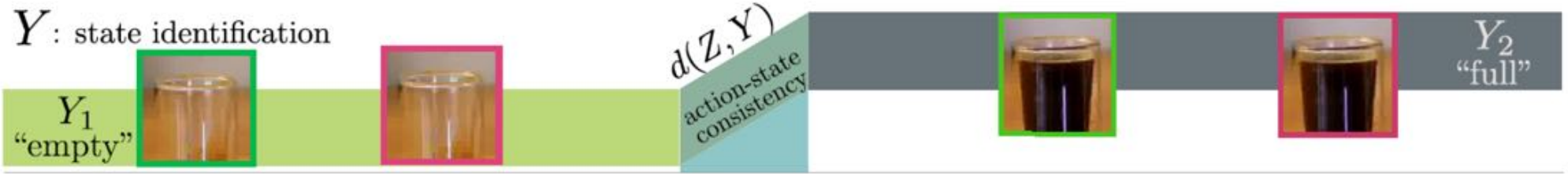
Action-state consistency

$$\underset{\substack{Y\in\{0,1\}^{M\times 2}\\ Z\in\{0,1\}^{T}}}{\text{minimize}} \quad f(Z) + g(Y) + d(Z,Y)$$

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

Action localization **Object state labeling**

Modelling Object States



Cost function: discriminative clustering

$$g(Y) = \min_{W_s \in \mathbb{R}^{d_s \times 2}} \underbrace{\frac{1}{2M}\|Y - X_s W_s\|_F^2}_{\text{Discriminative loss on data}} + \underbrace{\frac{\mu}{2}\|W_s\|_F^2}_{\text{Regularizer}}$$

OUTPUT (Discovered states)

Representation of tracklets (ROI pooling, ResNet 50) [Mxd$_s$] matrix

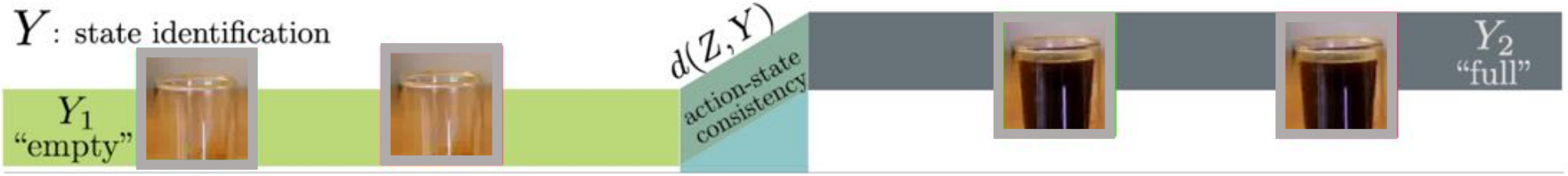
Linear state classifier [dx2] matrix

Constraints: $Y \in \mathcal{Y}$

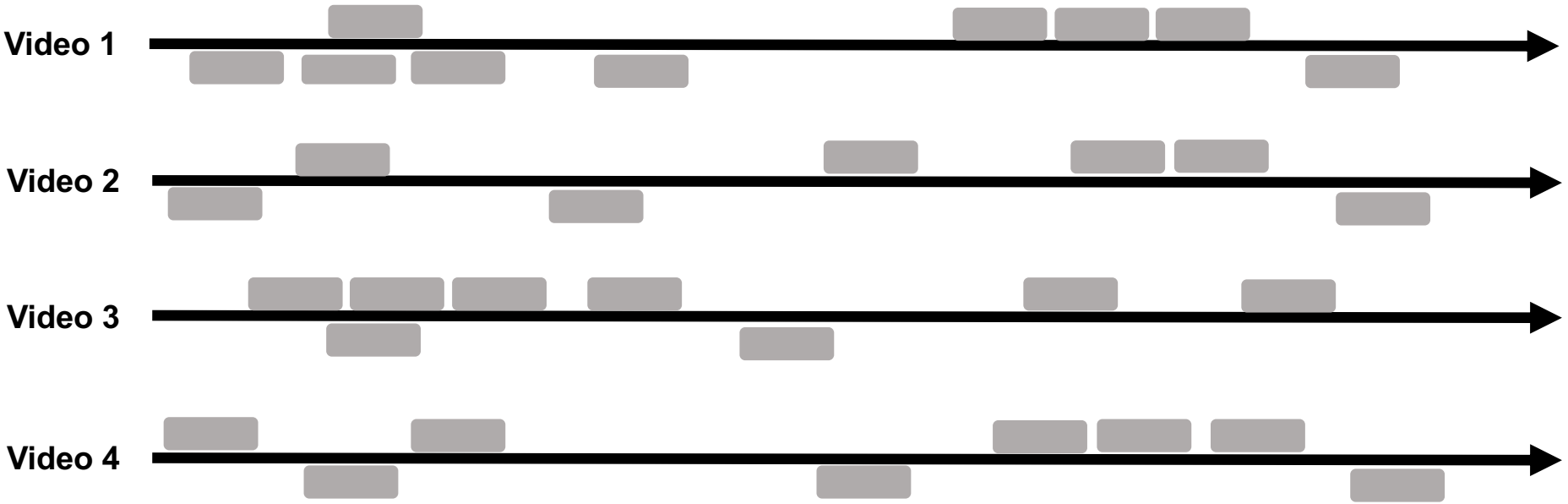
- **``Non-overlap'':** Only one object is manipulated at a time
- **Ordering constraints:** State 1 → State 2
- **Find at least one tracklet for each state**

See also: [Bach and Harchoui'07, Xu et al.'04, Doersch et al.'12, Joulin et al.'14, Bojanowski et al.'15, Hariharan et al.'12, Gharbi et al.'12]

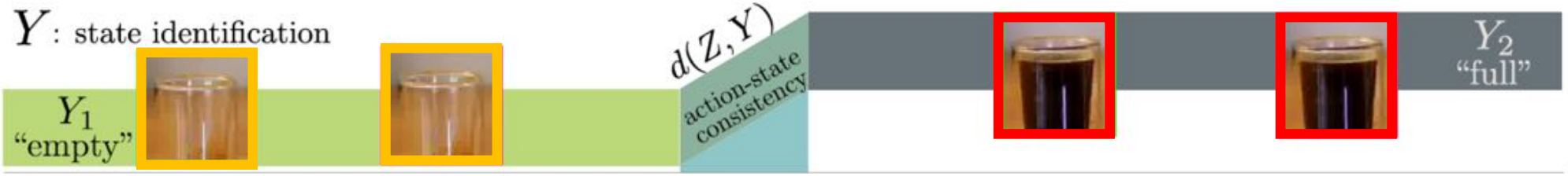
Modelling Object States



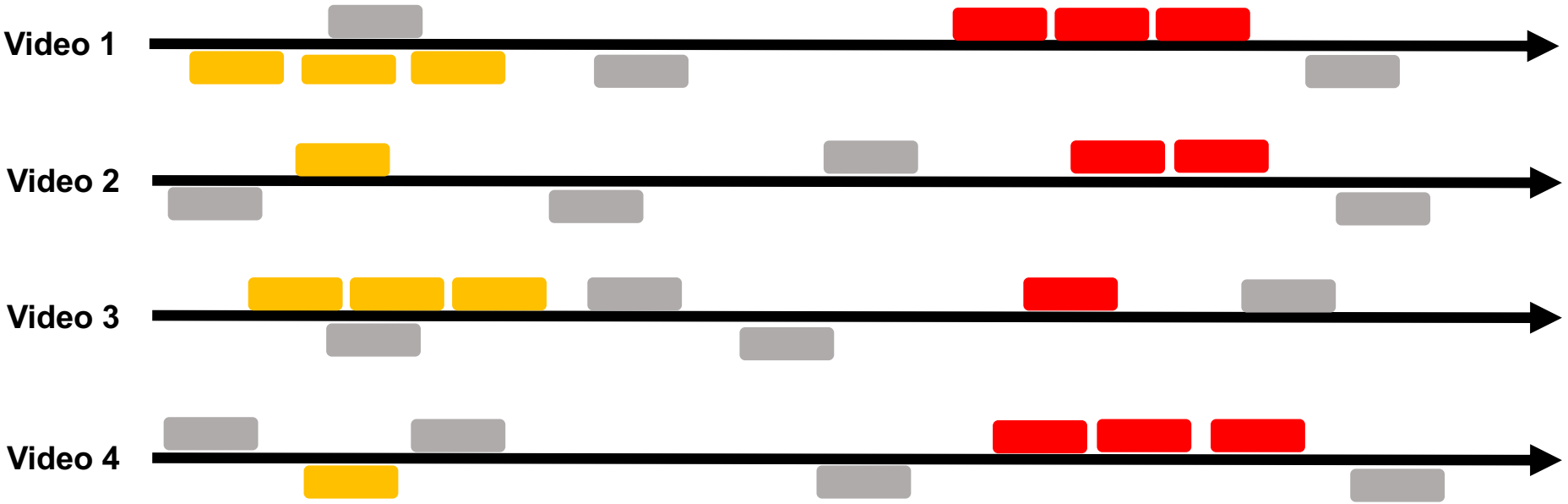
Cost function: discriminative clustering



Modelling Object States



Cost function: discriminative clustering



Modelling actions

$$f(Z) = \min_{W_v \in \mathbb{R}^{d_v}} \underbrace{\frac{1}{2T}\|Z - X_v W_v\|_F^2}_{\text{Discriminative loss on data}} + \underbrace{\frac{\lambda}{2}\|W_v\|_F^2}_{\text{Regularizer}},$$

OUTPUT (Time localization of action)

Representation of time intervals (IDT+CNN) [Txdv] matrix

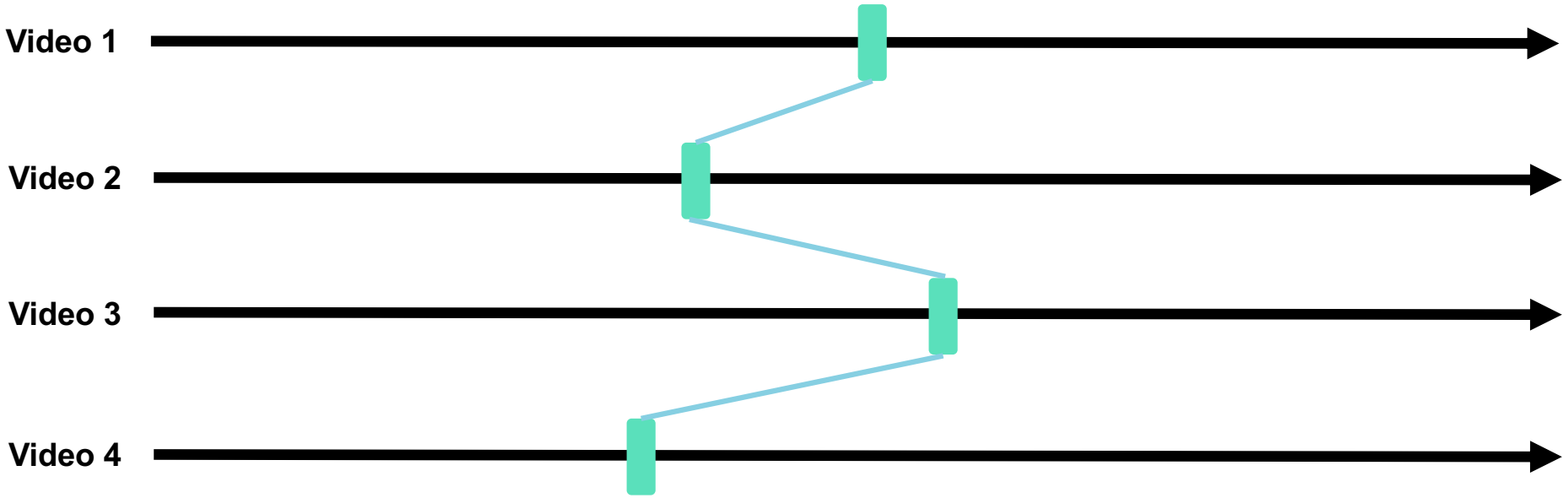
Linear action classifier [dvx1] matrix

Constraints: $Z \in \mathcal{Z}$

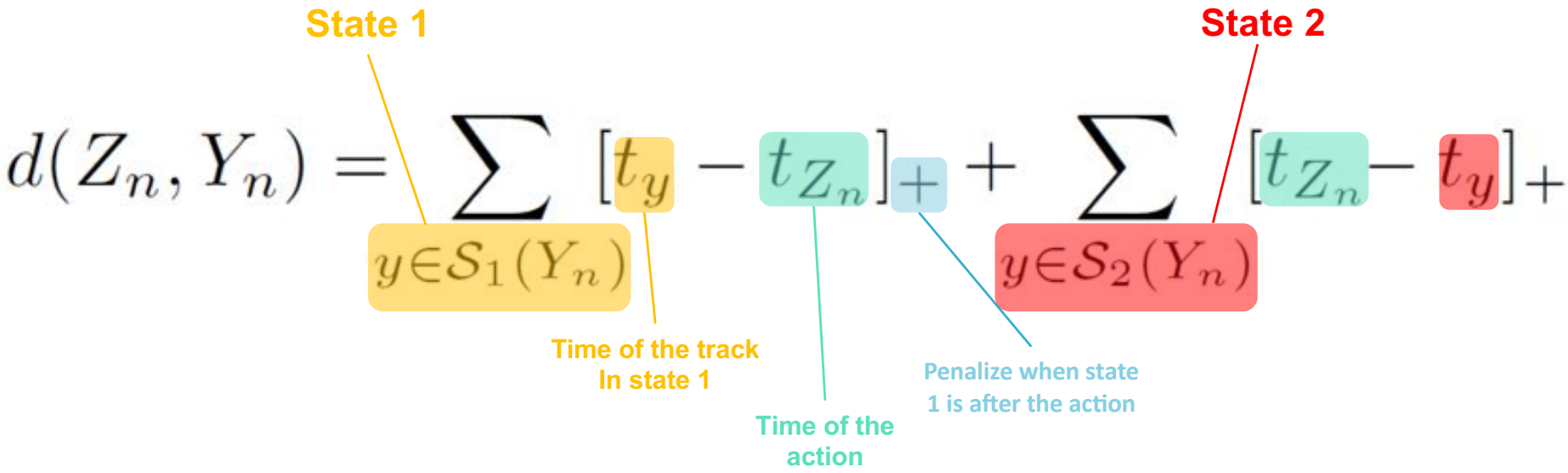
- **Saliency:** Select one time interval from each video

[Bojanowski et al, ICCV 2015]

Modelling actions



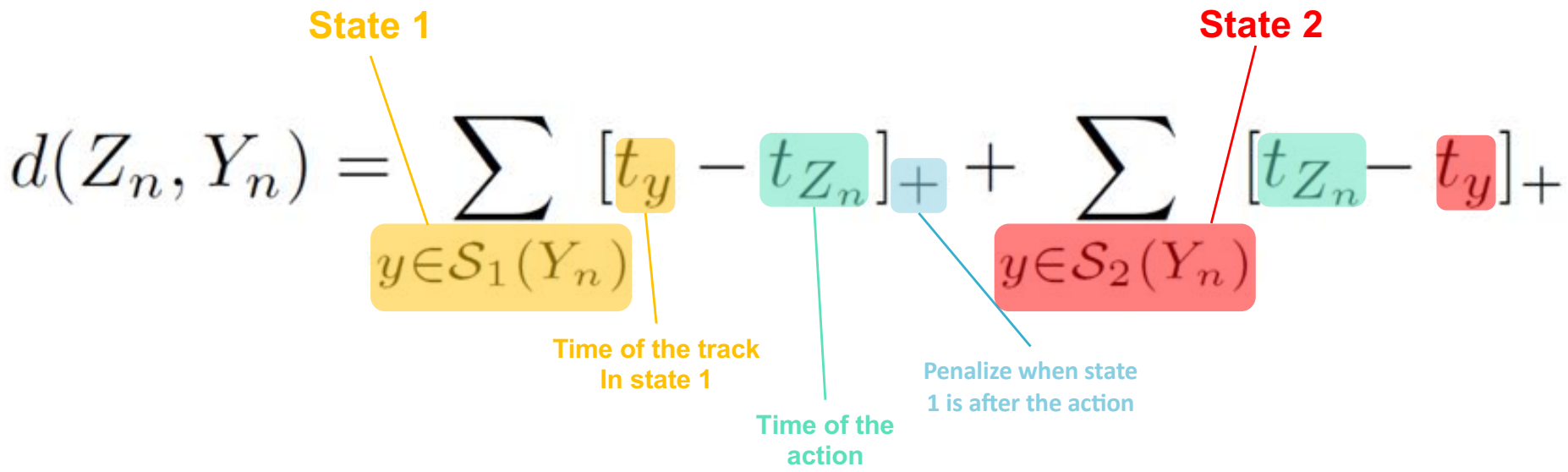
Joint cost function: Action-States

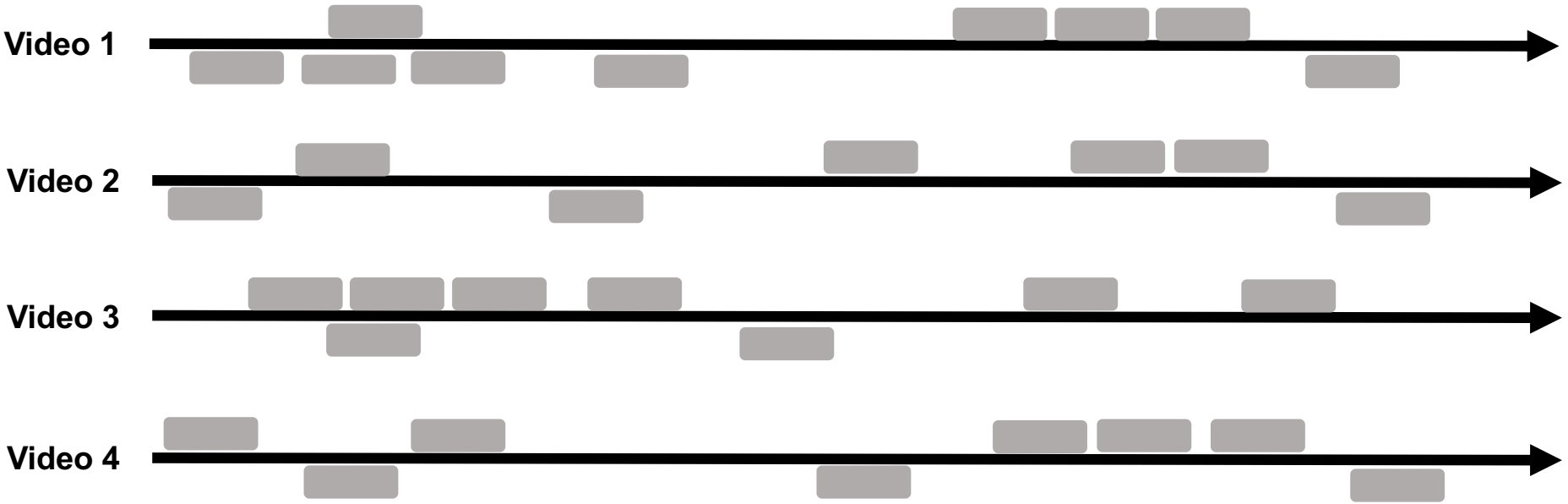


Main idea:

- **Action-State consistency:** action should be in between the initial and final state.

Joint cost function: Action-States





Joint cost function: Action-States

State 1

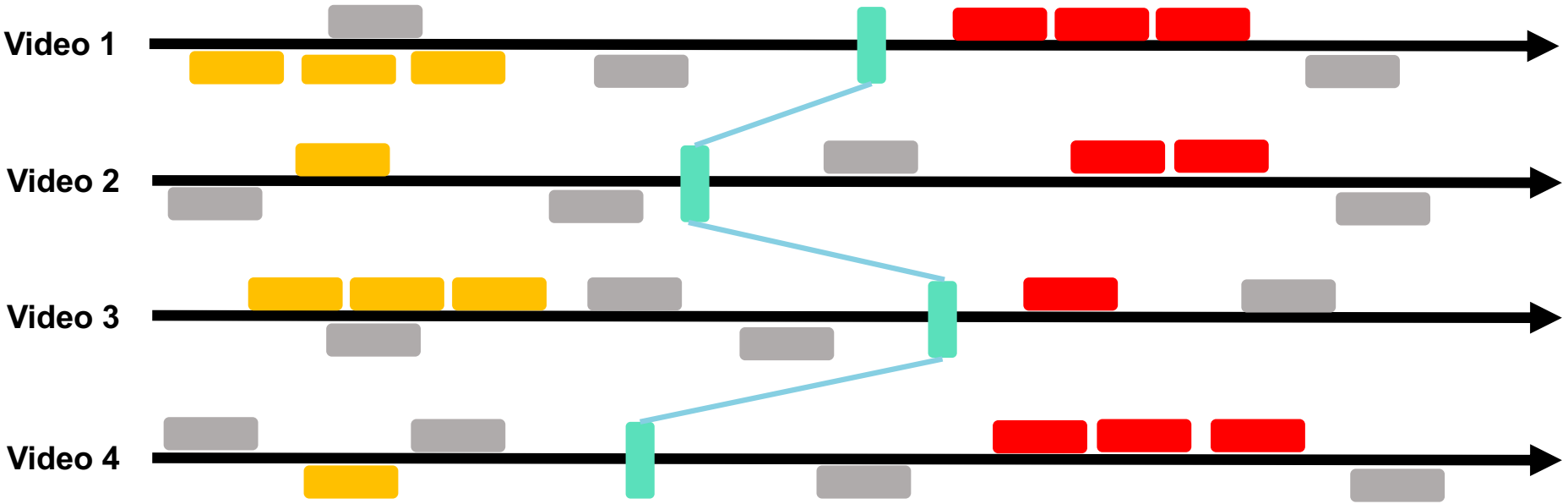
State 2

$$d(Z_n, Y_n) = \sum_{y \in \mathcal{S}_1(Y_n)} [t_y - t_{Z_n}]_+ + \sum_{y \in \mathcal{S}_2(Y_n)} [t_{Z_n} - t_y]_+$$

Time of the track In state 1

Time of the action

Penalize when state 1 is after the action



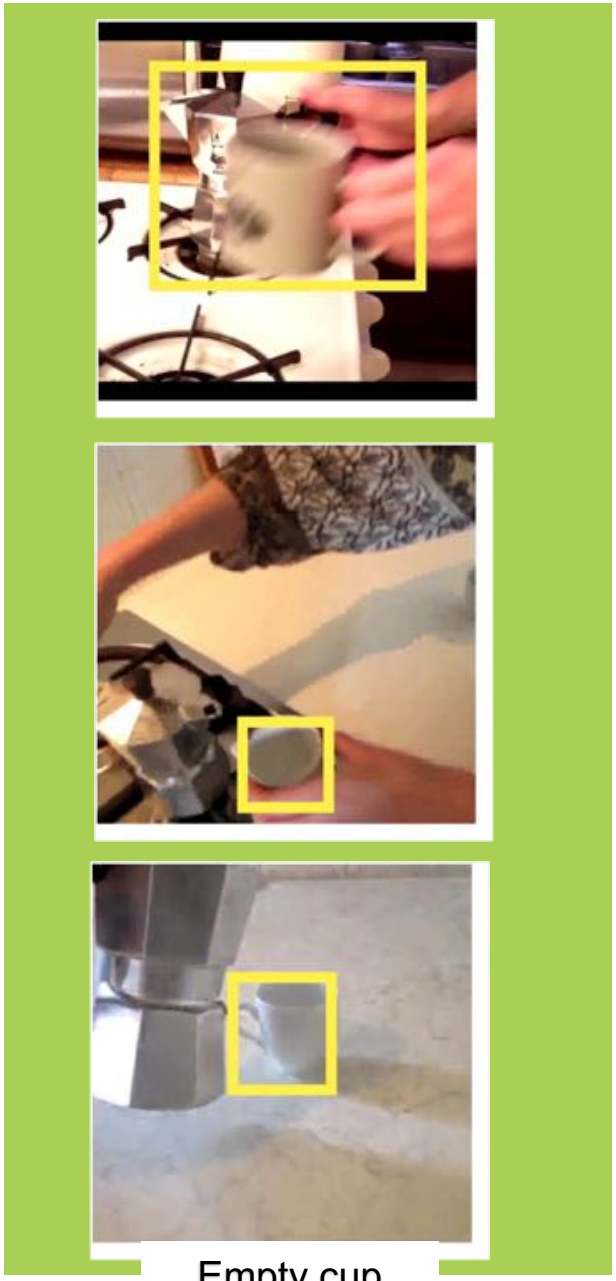
Dataset

- **7** actions
- ~**20-30s** per video
- Time annotation for actions
- Track level annotation for states with the following labels: state 1, state 2, false positive and ambiguous.
- Video are extracted from YouTube, Instruction videos [CVPR2016] and from the Charades dataset [ECCV2016]

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

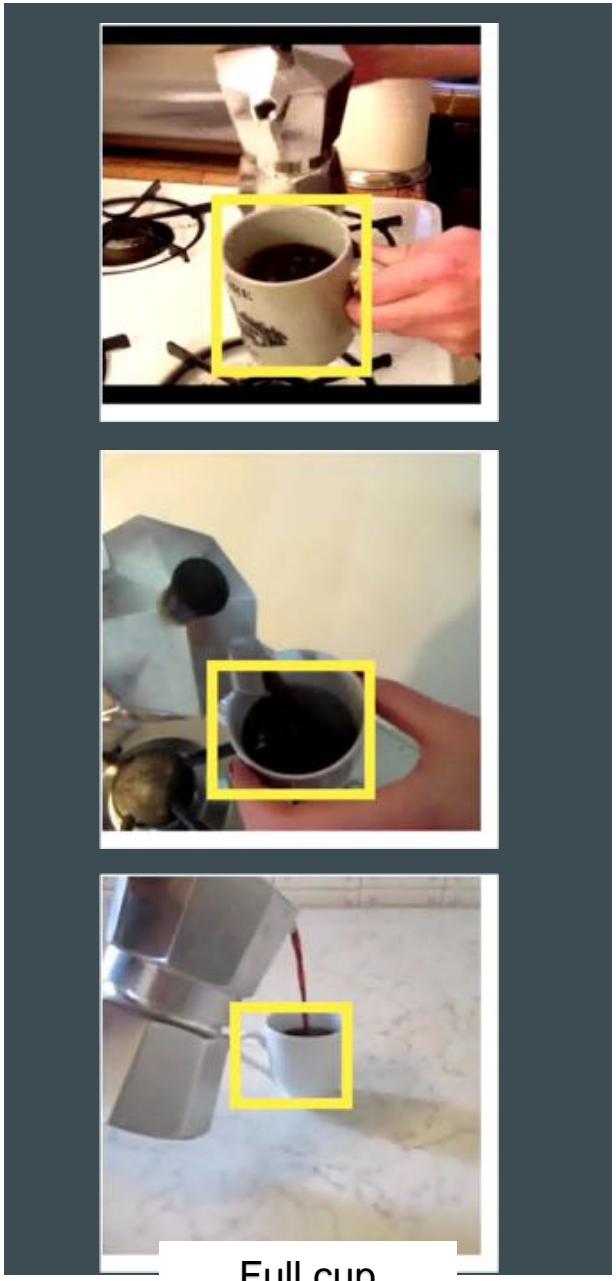
Qualitative results

Pour coffee



Empty cup

Pouring



Full cup

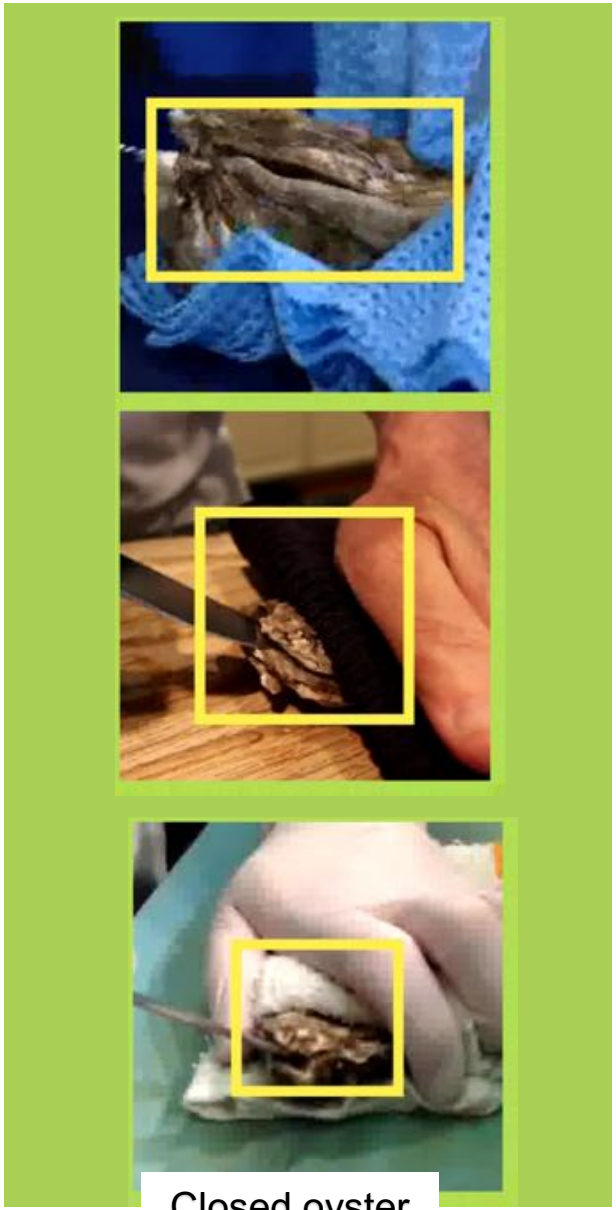
Put wheel

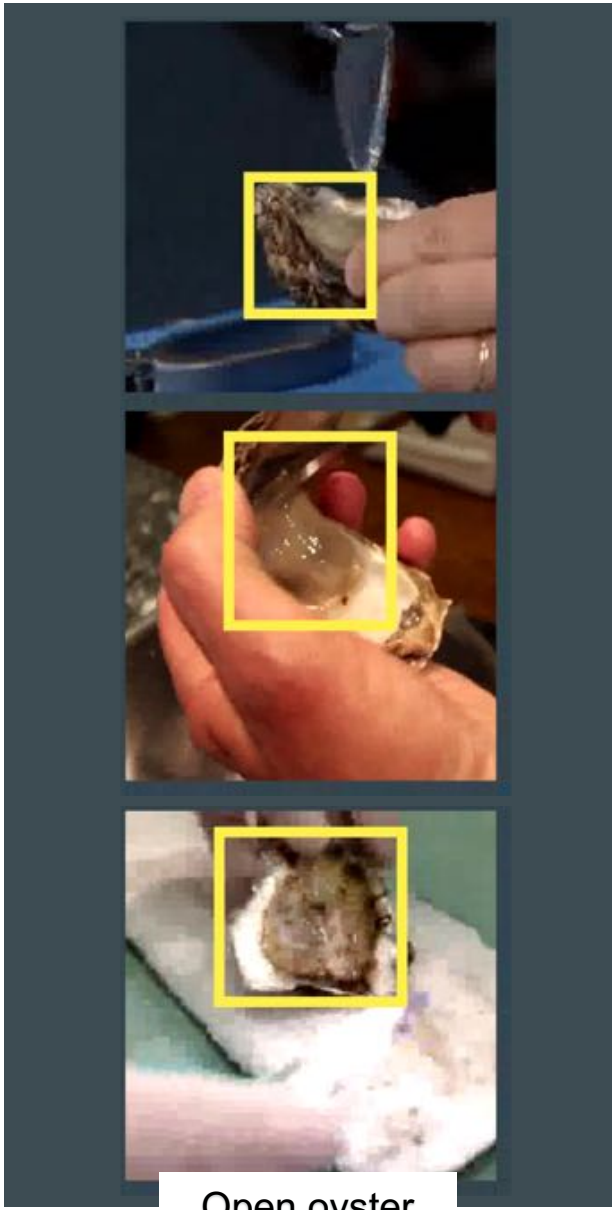
Wheel off car

Put wheel

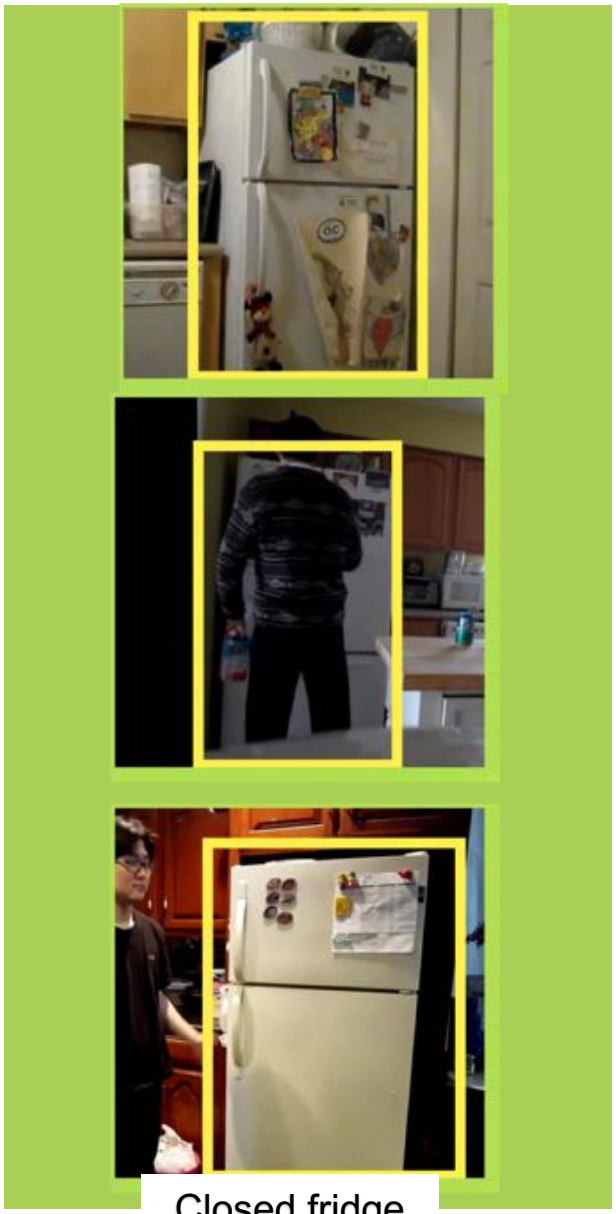
On car wheel

Open oyster



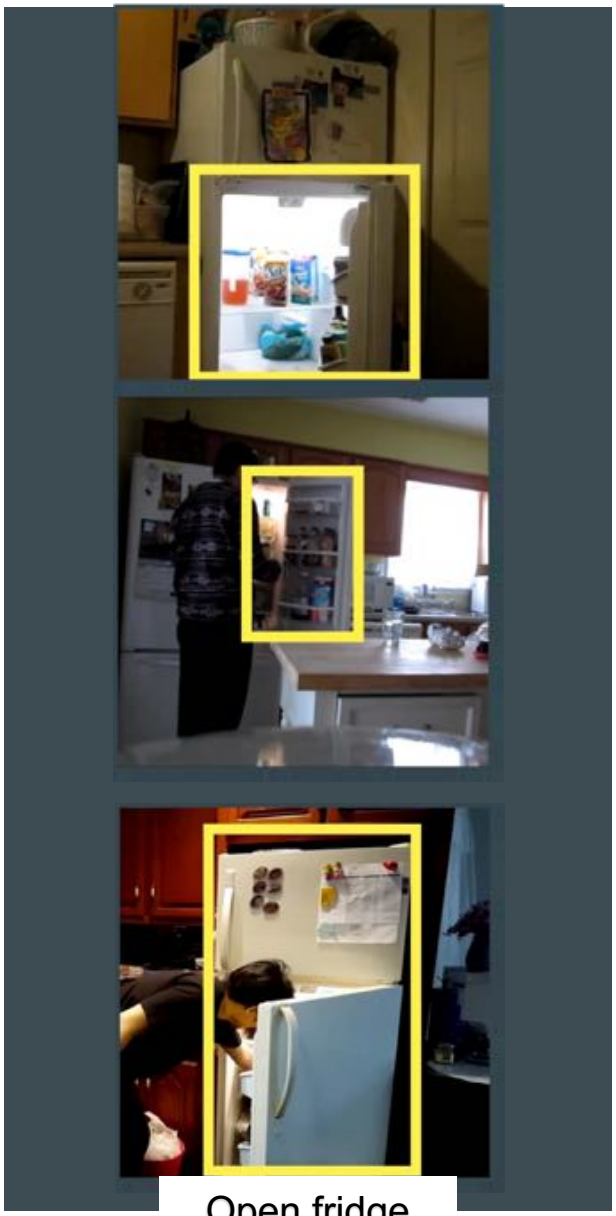


Open fridge



Closed fridge

Open



Open fridge

Quantitative results

Actions -> objects

Vs.

Objects -> actions

- Performance measured by percentage of tracklets with predicted state where the state is correct (**precision**)
- **Fixed minimal recall**: forced to predict in each video

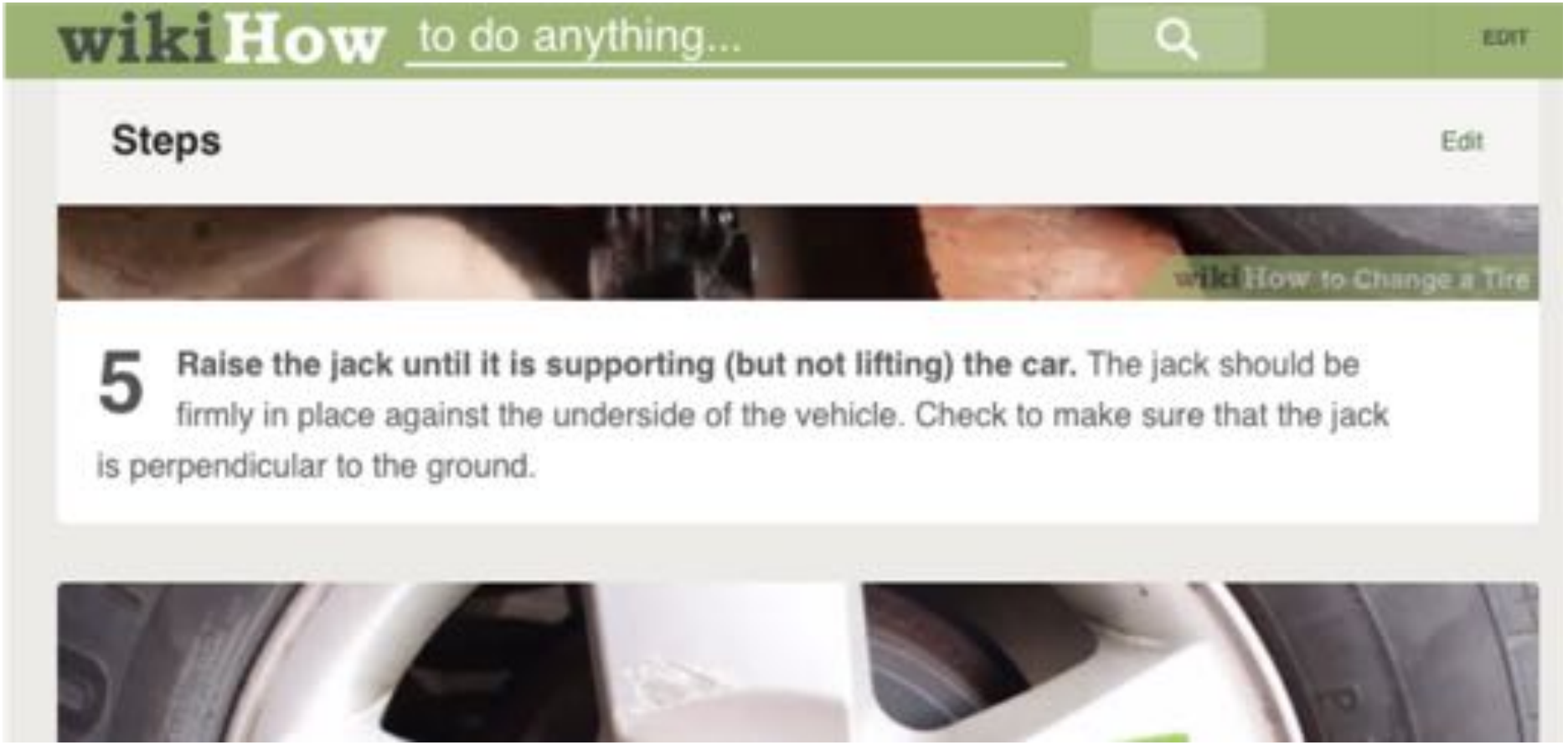
Results

	Method	put wheel	remove wheel	fill pot	open oyster	fill coff.cup	open fridge	close fridge	**Average**
State discovery	**(a)** Constraints only	0.35	0.38	0.35	0.36	0.31	0.29	0.42	0.35
	(b) Kmeans	0.25	0.12	0.11	0.23	0.14	0.19	0.22	0.18
	(c) Salient state only	0.33	0.60	0.19	0.25	0.14	0.43	0.39	0.33
	(d) At least one state only	**0.51**	**0.65**	0.33	0.48	0.28	0.45	0.35	0.44
	(e) Joint model	0.42	0.63	0.48	0.59	0.24	**0.50**	**0.49**	0.48
	(f) Joint model + det. scores.	0.47	**0.65**	**0.50**	**0.61**	**0.44**	0.46	0.43	**0.51**
	(g) Joint + GT act. feat.	0.59	0.63	0.56	0.50	0.32	0.51	0.50	0.52
Action localization	**(i)** Chance	0.31	0.20	0.15	0.11	0.40	0.23	0.17	0.22
	(ii) [5]	0.22	0.13	0.15	0.07	0.33	0.35	0.21	0.21
	(iii) [5] + object cues	0.26	0.17	0.15	0.14	**0.84**	0.34	0.36	0.32
	(iv) Joint model	**0.57**	**0.58**	**0.33**	**0.32**	0.83	**0.48**	**0.37**	**0.50**
	(v) Joint + GT stat. feat.	0.72	0.66	0.41	0.43	0.84	0.50	0.45	0.57

- Actions -> Objects: (d) 0.44 -> (e) 0.48
- **Objects -> Actions: (ii) 0.21 -> (iv) 0.50**

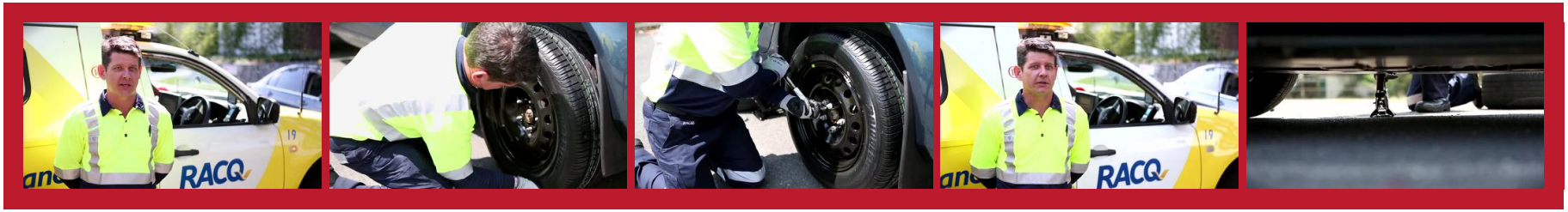
Weakly supervised object state learning

- Train text SVM from (~12) positive and (~50) negative training sentences from **Wikihow**
- Localize 20s candidate clips using transcribed narration



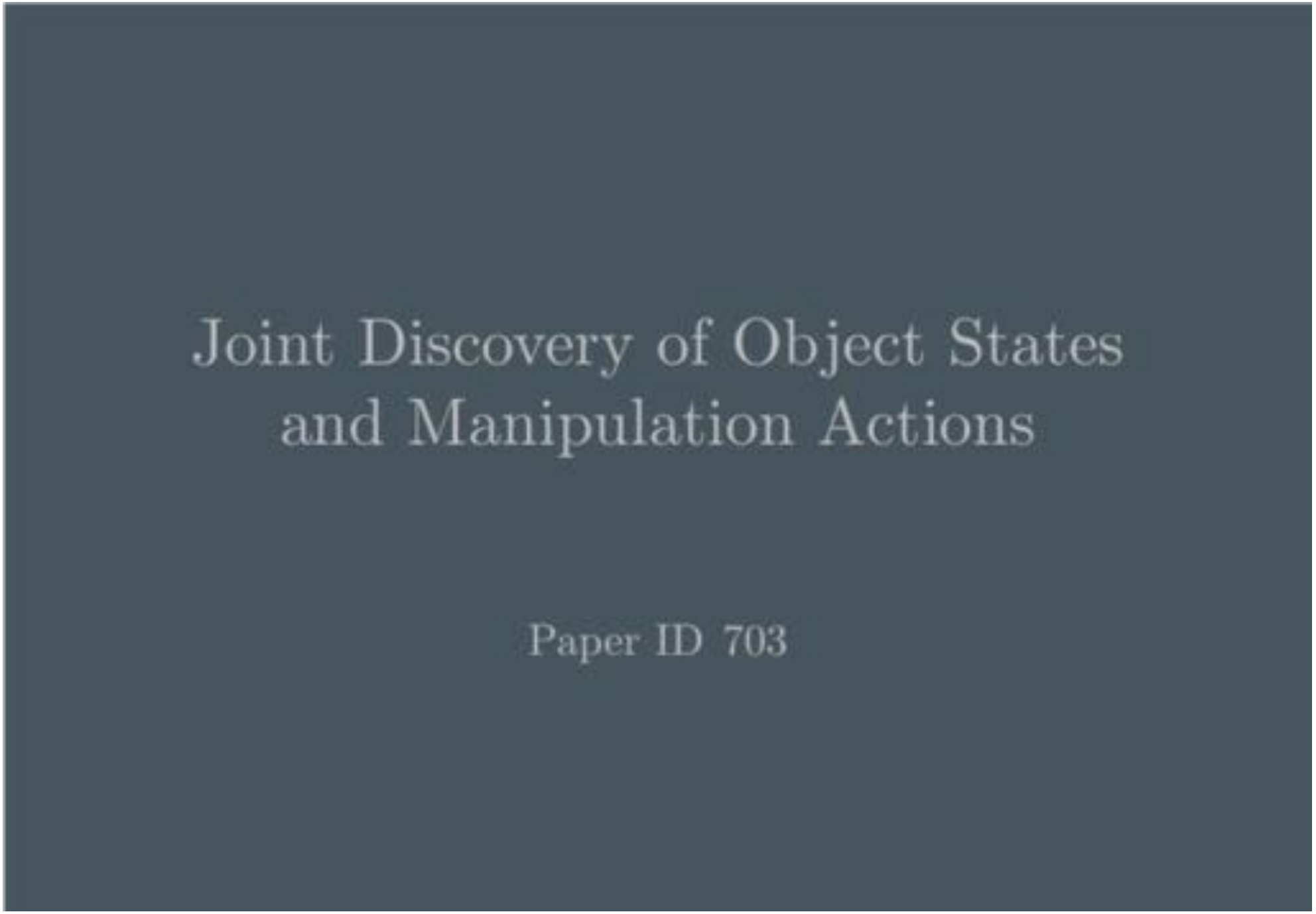
Weakly supervised object state learning

does vary from car to car so check your owner 's manual and that 'll identify exactly where in your particular car those items are kept what you 'll need is a wheel brace to enable you to undo the wheel nuts you 'll also need a jack to raise the vehicle to allow you to get the flat tyre off you start by removing the hub cap or the wheel nut covers whatever 's fitted to your vehicle , then using the wheel brace you want to loosen each of the wheel nuts half to approximately one turn before jacking up the vehicle next , we 're going to raise the vehicle to allow us to remove the flat tyre the jacking point and the type of jack used does vary from car to car so again its worth checking your owners manual . once the jack 's in place you then commence jacking the vehicle to release the weight of the vehicle from the flat tyre . now that we 've got the wheel off the ground we can remove each of the wheel nuts until all 5 or 6 or 4 , depending on whatever your car 's fitted with are off and now we can remove the flat tyre . a good idea is to place the flat wheel under the vehicle just for a safety feature in case the jack were to give way you



've then got something to support the vehicle so lift the spare wheel up onto the hub and make sure it 's sitting nice and flat and square to the vehicle then start each of the wheel nuts by hand then remove the flat tyre from under the vehicle and lower the jack . lower the vehicle down until the weight to the vehicle is on the spare tyre that 's just been fitted we 'll then use the wheel brace to tighten each of the wheel nuts in sequence . if you ca n't get to a tyre repairer in the short term it 's worth checking those wheel nuts again after a couple hundred kilometres of driving , just to make sure they are still tight i guess the main thing is to know your vehicle , so you should be checking that your spare tyre is fitted to the vehicle and inflated and also that all of your tools are in a serviceable condition . with some vehicles it can vary , you 've got emergency spares or space savers as they are commonly called and they 're only for literally emergency purposes to get you to a place of repair as quickly as possible when in doubt ring

Qualitative results



Quantitative results

	Method	put wheel	remove wheel	fill pot	open oyster	fill coff.cup	Ave.
State disc.	**(a)** Chance	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
	(f) Joint	**0.38**	**0.53**	0.25	**0.43**	**0.20**	**0.36**
	(f) 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

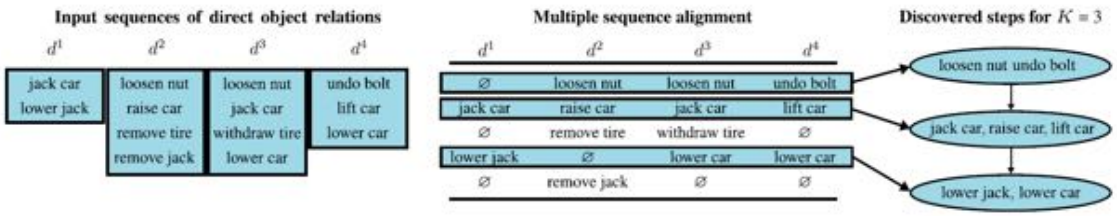
- Objects -> Actions vs. Actions -> Objects
- Joint method performs best in both cases.

Outline

1. Learn sequence of main steps of a task

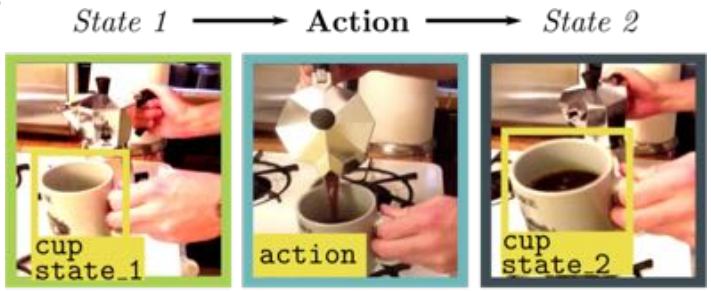
   [Alayrac et al., CVPR 2016]

   [Alayrac et al., PAMI 2017]

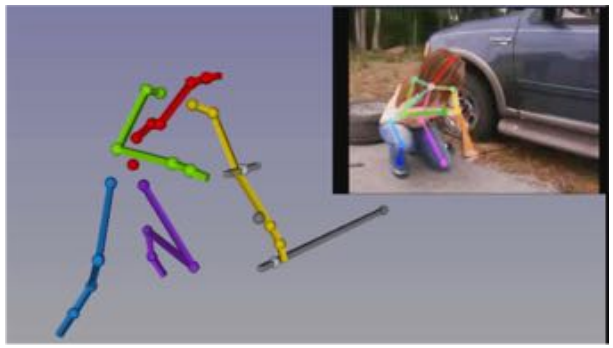


2. Modeling changes in object states

   [Alayrac et al., ICCV 2017, to appear]



3. Discussion and challenges



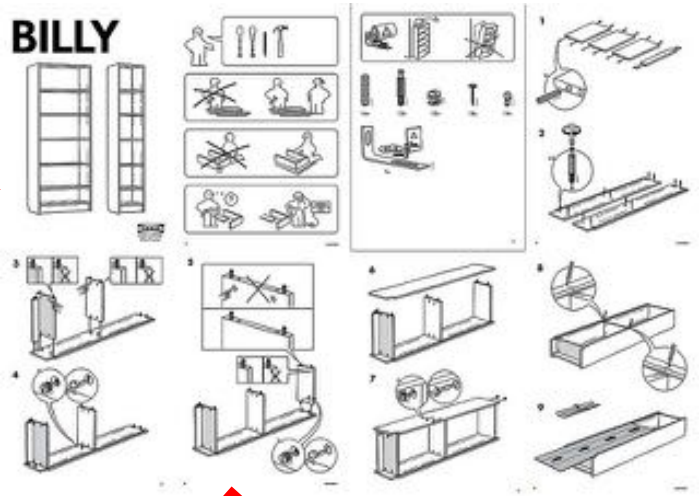
Challenge I.: Learning constraints

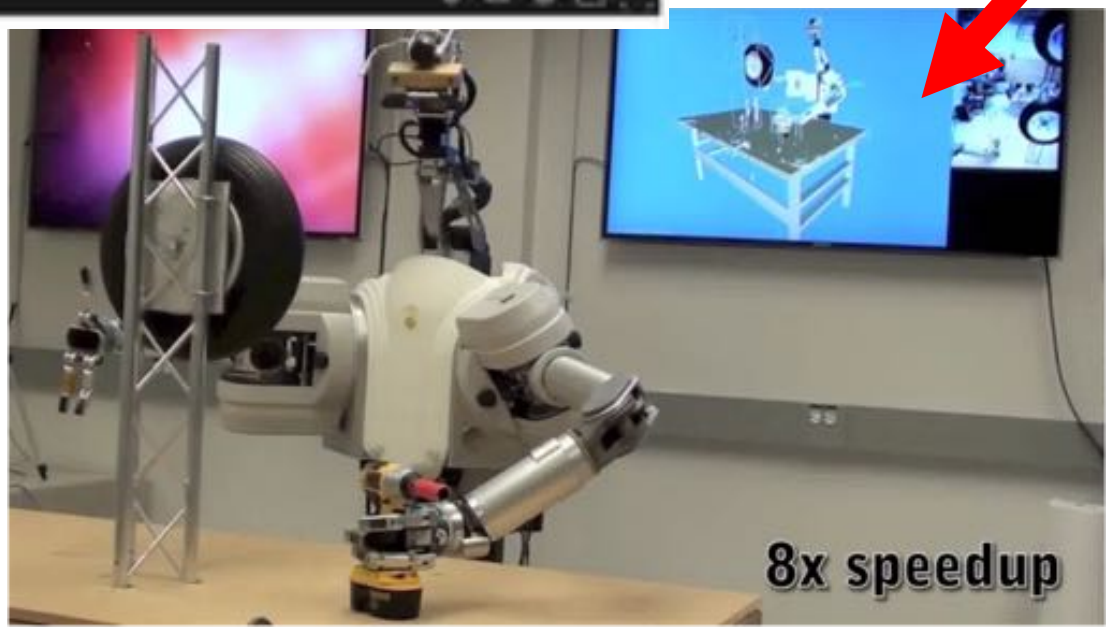
- **Learning visual representations** with constraints from
  - **Language** (narration)

  - **Physical structure** (plausible temporal and spatial relations)

- **What constraints** for unsupervised learning?
- Can constraints be **learnt from data?**

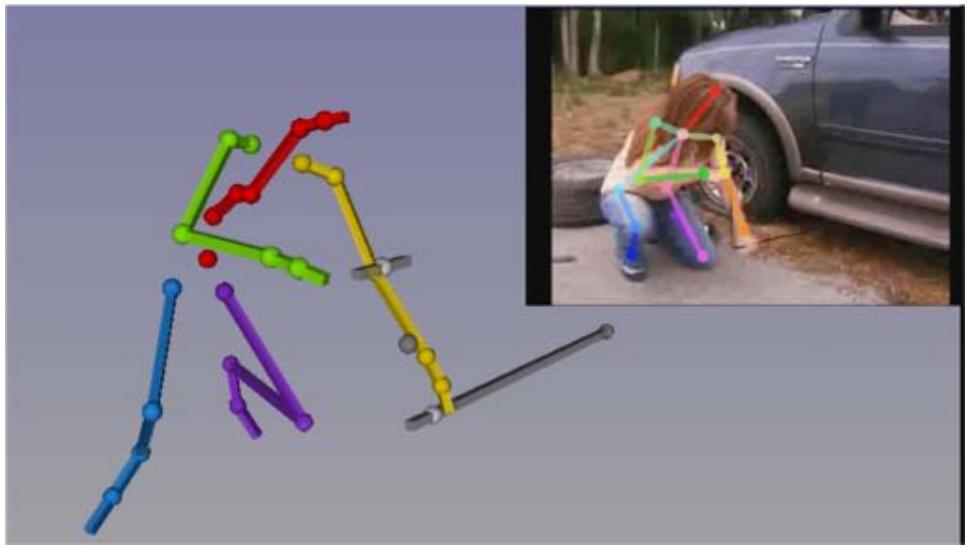
Challenge II.: How to generalize to new conditions?

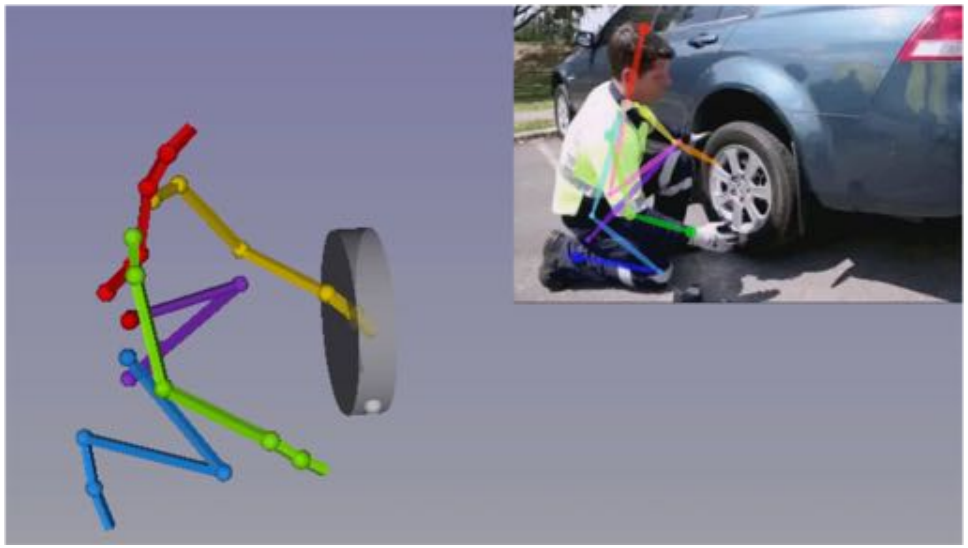




[Hacket et al., 2013]

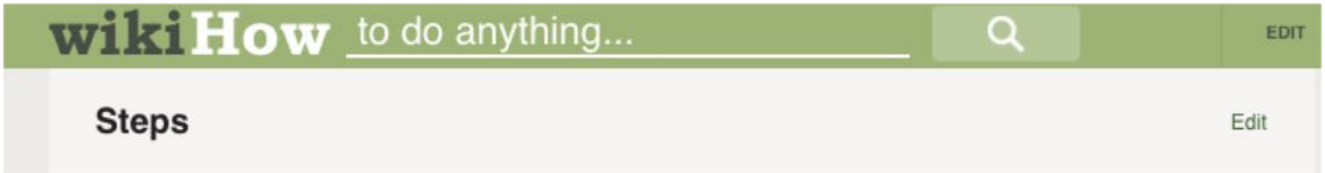
Example : Learn manipulation from video





[Li, Mansard, Laptev, Sivic, 2017], See also: [Bogo et al., SMPL, ECCV 2016], [Pischulin et al., DeepCut, CVPR 2016], [Carpentier, Valenza, Mansard, et al. 2015]

Challenge III: How can we learn from one example?



Scaling-up current dataset to 1000 of tasks and 100,000s of videos.