Unsupervised Learning from Narrated Instruction Videos

Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, Simon Lacoste-Julien

INRIA/ENS, Paris, France

Goal and overview

The problem:

Automatically learn the **main steps** to complete a given **task** from **narrated instruction videos**.

Input: A set of narrated instruction videos.

Start by loosening each bolt. Then locate the jack and lift the car. Now you can remove the bolts and then the wheel.

First undo the nuts. Once that done, you can jack the car. Then withdraw the nuts completely so that you can remove the flat tire.

Outputs:

- **List of K (input) main steps**
- **Visual and linguistic representations of the steps**
- **Temporal localization of each step in the videos**

1. Loosen nuts
2. Jack the car
3. Remove the flat tire

Contributions:

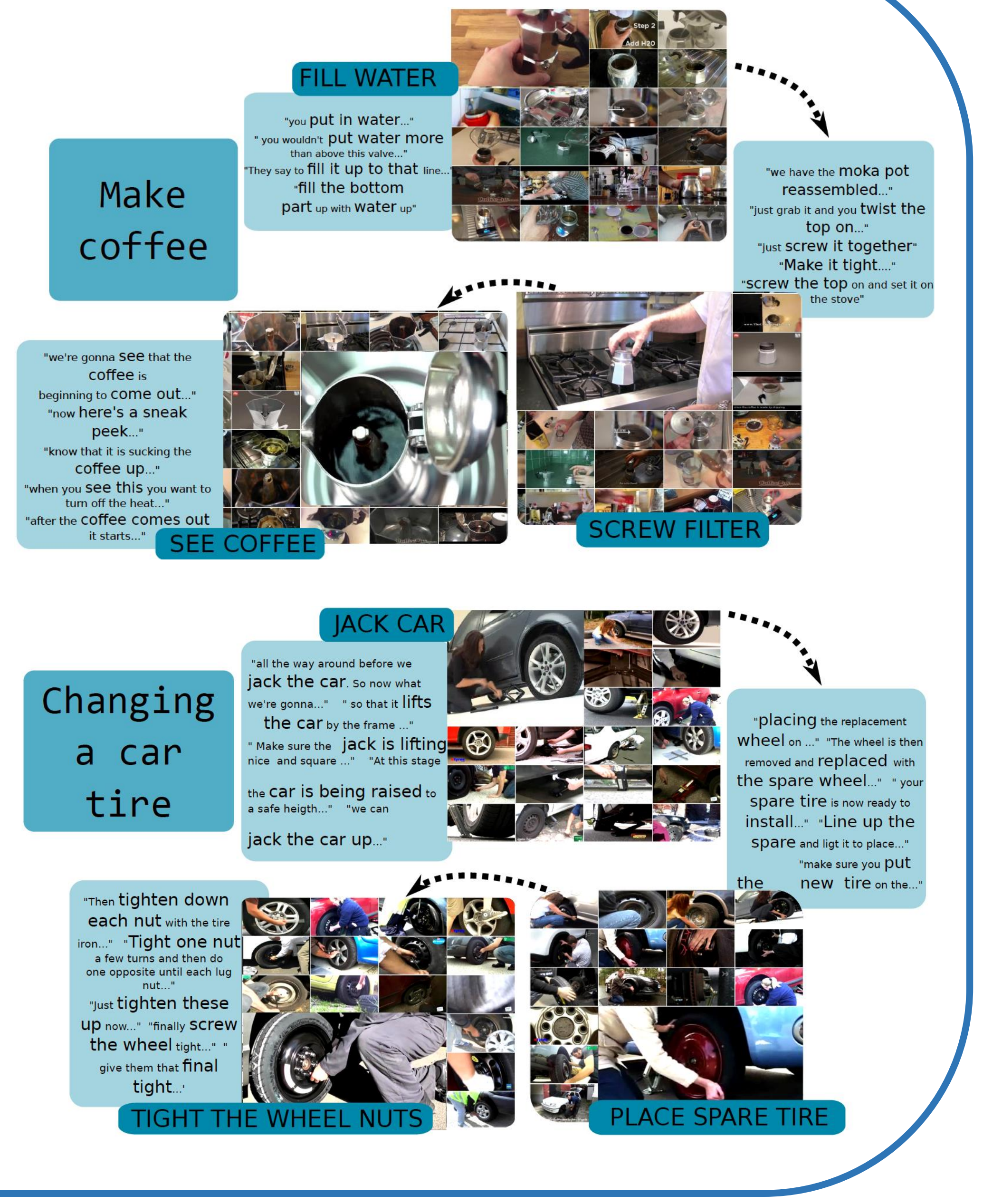
i. Collected and annotated a **new dataset** of narrated instruction videos,

ii. Developed an **unsupervised learning method** that takes advantage of the complementary nature of the text and video,

iii. Experimentally demonstrated **recovery of main steps** and their **locations in video**.

A new dataset

- 5 tasks:
  - ➢ Changing a car tire
  - ➢ Repot a plant
  - ➢ Make coffee
  - ➢ Perform CPR
  - ➢ Jump a car battery
- **30** videos per task (total of **800,000** frames).
- Manual correction of the ASR transcripts.
- **Manual annotation** of 7-10 main steps and time localization in each video (only used for evaluation).


Make coffee
FILL WATER
"you put in water..." " you wouldn't put water more than above this valve..." "They say to fill it up to that line..." "fill the bottom part up with water up"
"we have the moka pot reassembled..." "just grab it and you twist the top on..." "just screw it together" "Make it tight...." "screw the top on and set it on the stove"
SCREW FILTER
"we're gonna see that the coffee is beginning to come out..." "now here's a sneak peek..." "know that it is sucking the coffee up..." "when you see this you want to turn off the heat..." "after the coffee comes out it starts..."
SEE COFFEE
Changing a car tire
JACK CAR
"all the way around before we jack the car. So now what we're gonna..." " so that it lifts the car by the frame ..." " Make sure the jack is lifting nice and square ..." "At this stage the car is being raised to a safe heigth..." "we can jack the car up..."
"placing the replacement wheel on ..." "The wheel is then removed and replaced with the spare wheel..." " your spare tire is now ready to install..." "Line up the spare and ligt it to place..." "make sure you put the new tire on the..."
PLACE SPARE TIRE
"Then tighten down each nut with the tire iron..." "Tight one nut a few turns and then do one opposite until each lug nut..." "Just tighten these up now..." "finally screw the wheel tight..." " give them that final tight...'
TIGHT THE WHEEL NUTS

Approach

Assumptions

- Each task is performed by an **ordered** sequence of steps.
- People do **what** they say (roughly) **when** they say it

Two linked clustering steps

1. Text clustering : multiple sequence **alignment**
2. Discriminative video clustering under **text constraints**

1st STEP : TEXT
Identify main steps

Multiple sequence alignment (MSA)

- Text signals are first processed into sequences of **direct object relations**: *Let's now jack the car.* ⟶ **DOBJ = (jack,car)**
- Similarity between **dobj** is obtained using **Wordnet**.

INDIVIDUAL INPUT SEQUENCES:

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

OUTPUT OF THE MSA:

New! Formulate **MSA** as a **QP**, and approximately solve it with **Frank-Wolfe**

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

FINAL OUTPUT:

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

Text assignment [SxK] matrix

2nd STEP : VIDEO
Temporally localize each steps

Discriminative clustering under text constraints

CONSTRAINED OPTIMIZATION PROBLEM:

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

Discovered temporal localization [TxK] matrix

Subtitle Alignment [SxT] matrix

where $h(Z)$ is a **discriminative clustering** cost [1]:

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

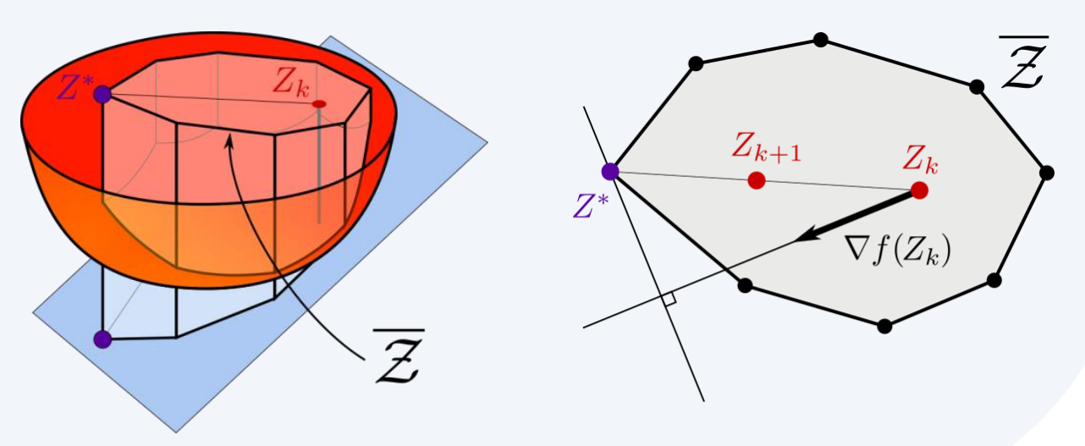
Representation of video (IDTF,CNN) [Txd] matrix

Linear action classifier [dxK] matrix

T = total number of time intervals
S = total number of text tokens
K = number of steps
d = dimension of the features

OPTIMIZATION METHOD [2,3]:

- Optimize convex relaxation using Frank-Wolfe
- Use **DP** as the linear oracle
- Cost classifier based **rounding**

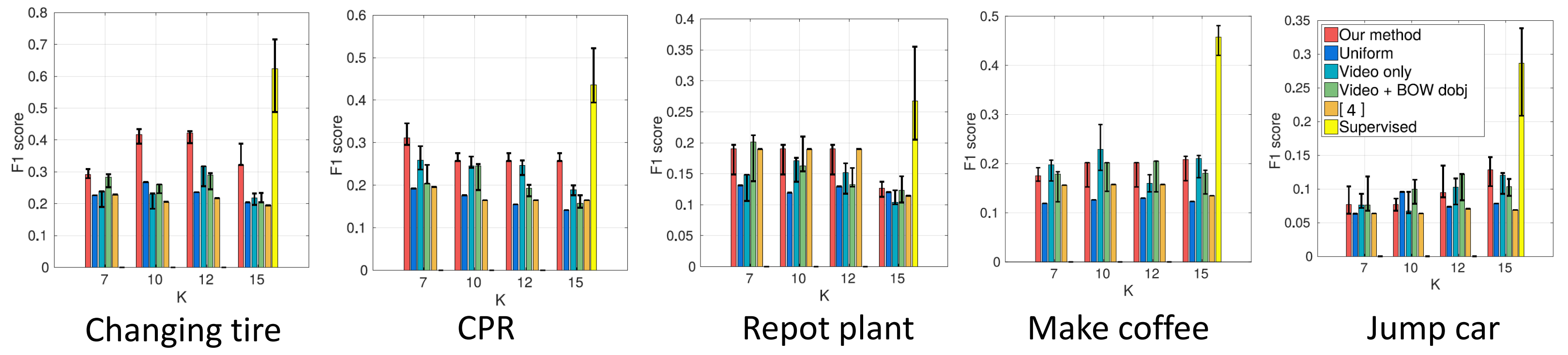


Experiments

Script discovery results

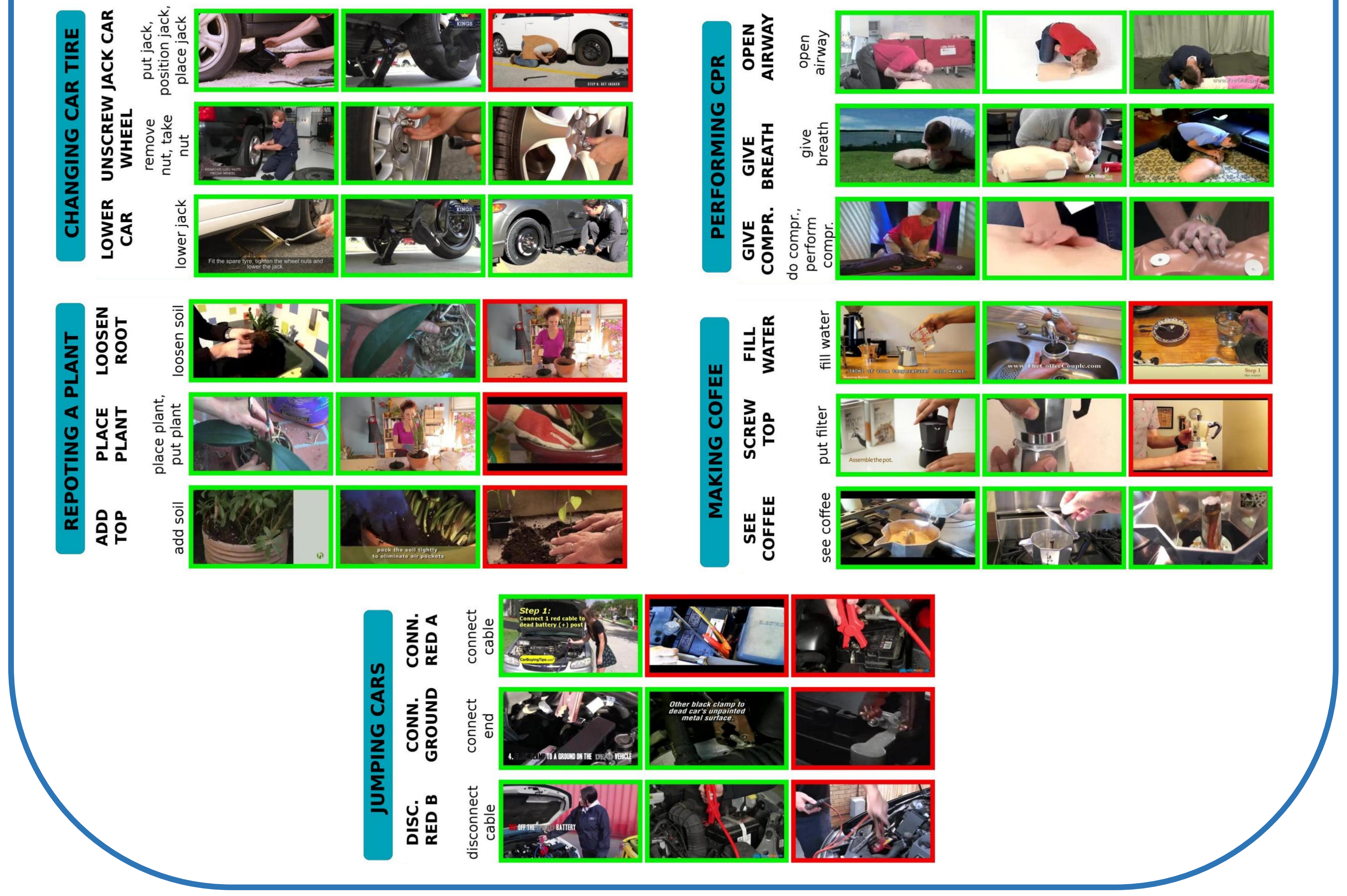
Changing a tire		Performing CPR		Repot a plant		Make coffee		Jump car	
GT (**11**)	$K \leq 10$	GT (**7**)	$K \leq 10$	GT (**7**)	$K \leq 10$	GT (**10**)	$K \leq 10$	GT (**12**)	$K \leq 10$
get tools out	**get tire**	*open airway*	**open airway**	*take plant*	**remove plant**	*add coffee*	**put coffee**	*connect red A*	**connect cable**
start loose	**loosen nut**	*check pulse*	**put hand**	*put soil*	**use soil**		**fill chamber**		**charge battery**
	put jack		**tilt head**	*loosen roots*	**loosen soil**	*fill water*	**fill water**	*connect red B*	**connect end**
jack car	**jack car**		**lift chin**	*place plant*	**place plant**	*screw filter*	**put filter**	*start car A*	**start car**
unscrew wheel	**remove nut**	*give breath*	**give breath**	*add top*	**add soil**		**see steam**	*remove cable A*	**remove cable**
remove wheel	**take wheel**	*do compressions*	**do compression**	*water plant*	**water plant**	*put stove*	**take minutes**	*remove cable B*	**disconnect cable**
put wheel	**take tire**		**open airway**				**make coffee**		
screw wheel	**put nut**		**start compression**			*see coffee*	**see coffee**		
lower car	**lower jack**		**do compression**			*pour coffee*	**make cup**		
tight wheel	**tighten nut**		**give breath**						
Precision	0.9	Precision	0.4	Precision	1	Precision	0.67	Precision	0.83
Recall	0.9	Recall	0.57	Recall	0.86	Recall	0.6	Recall	0.42

Localizing instruction steps in video



Changing tire   CPR   Repot plant   Make coffee   Jump car

Qualitative results



References

[1] Bach and Harchaoui. DIFFRAC: A discriminative and flexible framework for clustering. In *NIPS, 2007.*

[2] Bojanowski et al. Weakly supervised action labeling in videos under ordering constraints. In *ECCV, 2014.*

[3] Bojanowski et al. Weakly-Supervised Alignment of Video With Text. In *ECCV, 2015.*

[4] Malmaud et al. What's cookin'? Interpreting cooking videos using text, speech and vision. In *NACL, 2015.*

Check out our project webpage for code/data!