Learning from Video and Text via Large-Scale Discriminative Clustering

Antoine Miech[1,2], Jean-Baptiste Alayrac[1,2], Piotr Bojanowski[1], Ivan Laptev[1,2], Josef Sivic[1,2,3]

[1]Inria, Paris

[2]DI ENS, École Normale Supérieure,
PSL Research University, Paris

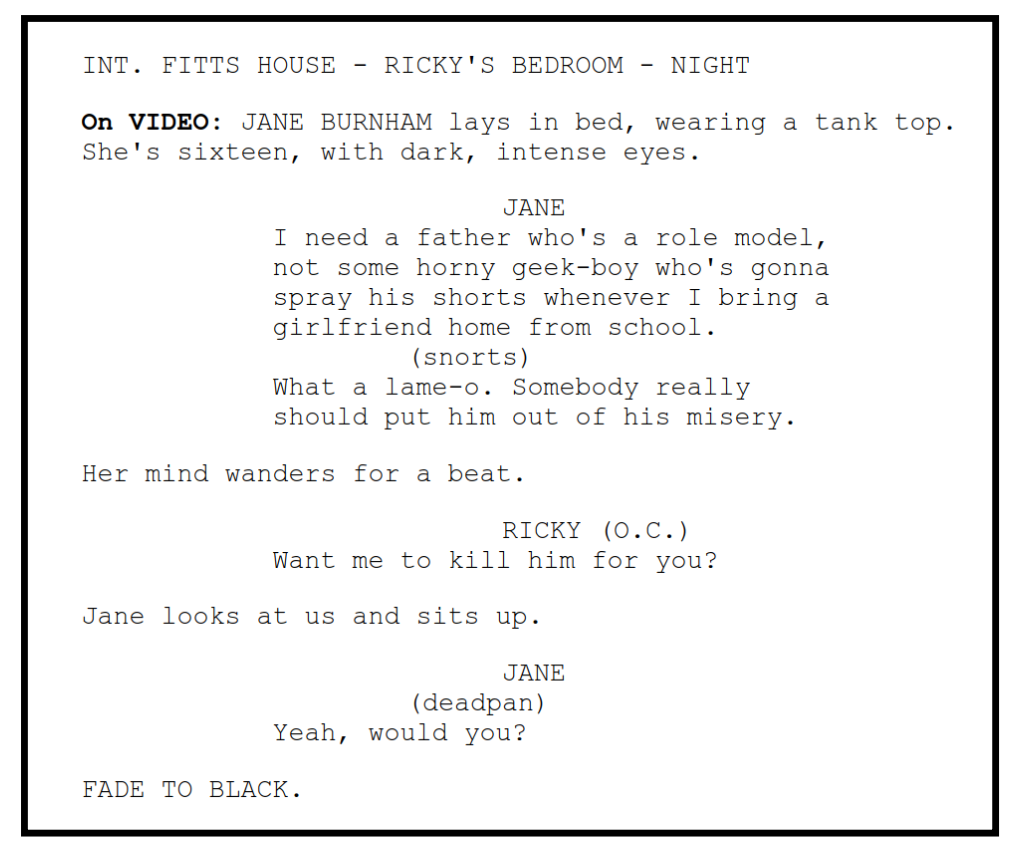
[3]CIIRC, CTU, Prague

Keywords: Text-Video, Weak-Supervision, Discriminative Clustering
Person-Action Recognition, Block-Coordinate Frank-Wolfe

Goal

- **Scale-up** discriminative clustering for weakly supervised learning
- Demonstrate **weakly supervised** learning of actors and actions on large-scale dataset of movies

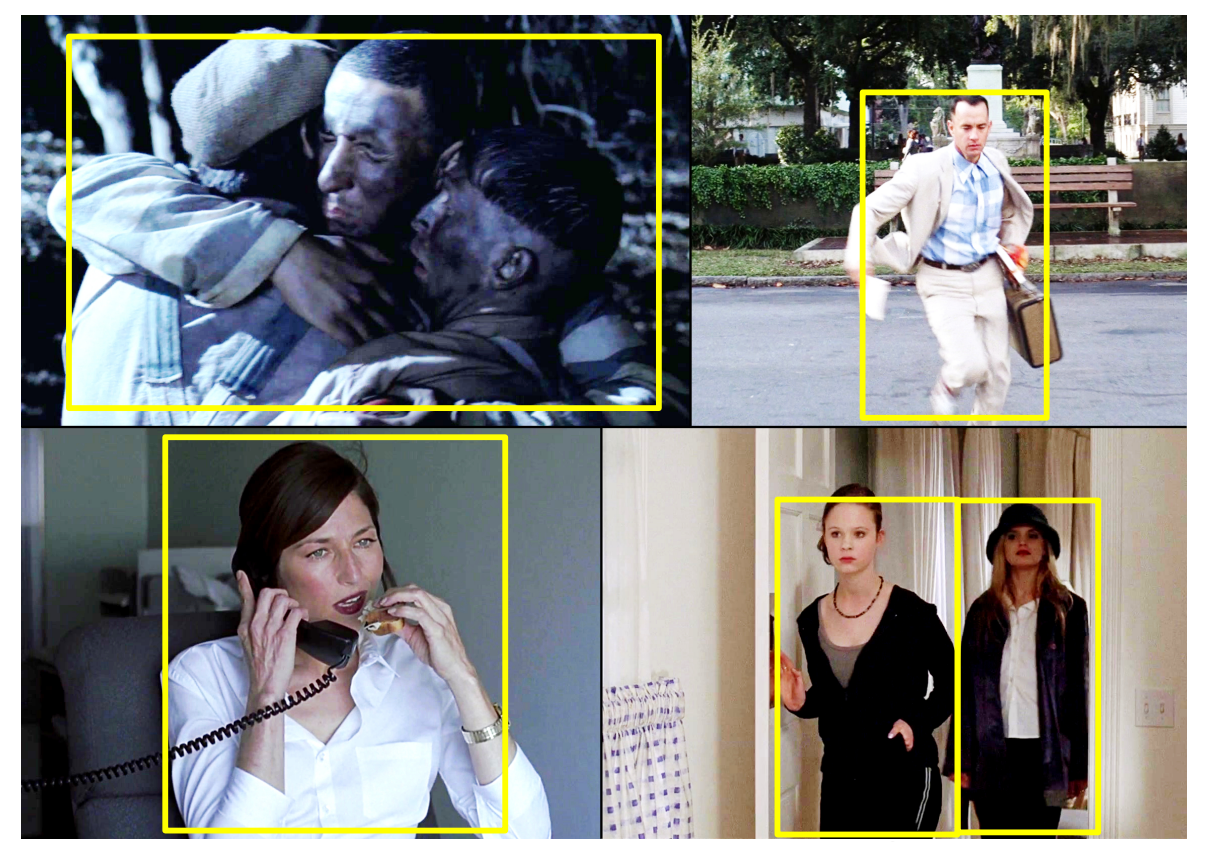
INPUT



```
INT. FITTS HOUSE - RICKY'S BEDROOM - NIGHT

On VIDEO: JANE BURNHAM lays in bed, wearing a tank top.
She's sixteen, with dark, intense eyes.

                      JANE
            I need a father who's a role model,
            not some horny geek-boy who's gonna
            spray his shorts whenever I bring a
            girlfriend home from school.
                   (snorts)
            What a lame-o. Somebody really
            should put him out of his misery.

Her mind wanders for a beat.

                      RICKY (O.C.)
            Want me to kill him for you?

Jane looks at us and sits up.

                      JANE
                   (deadpan)
            Yeah, would you?

FADE TO BLACK.
```

Movie script
"Free Annotation"

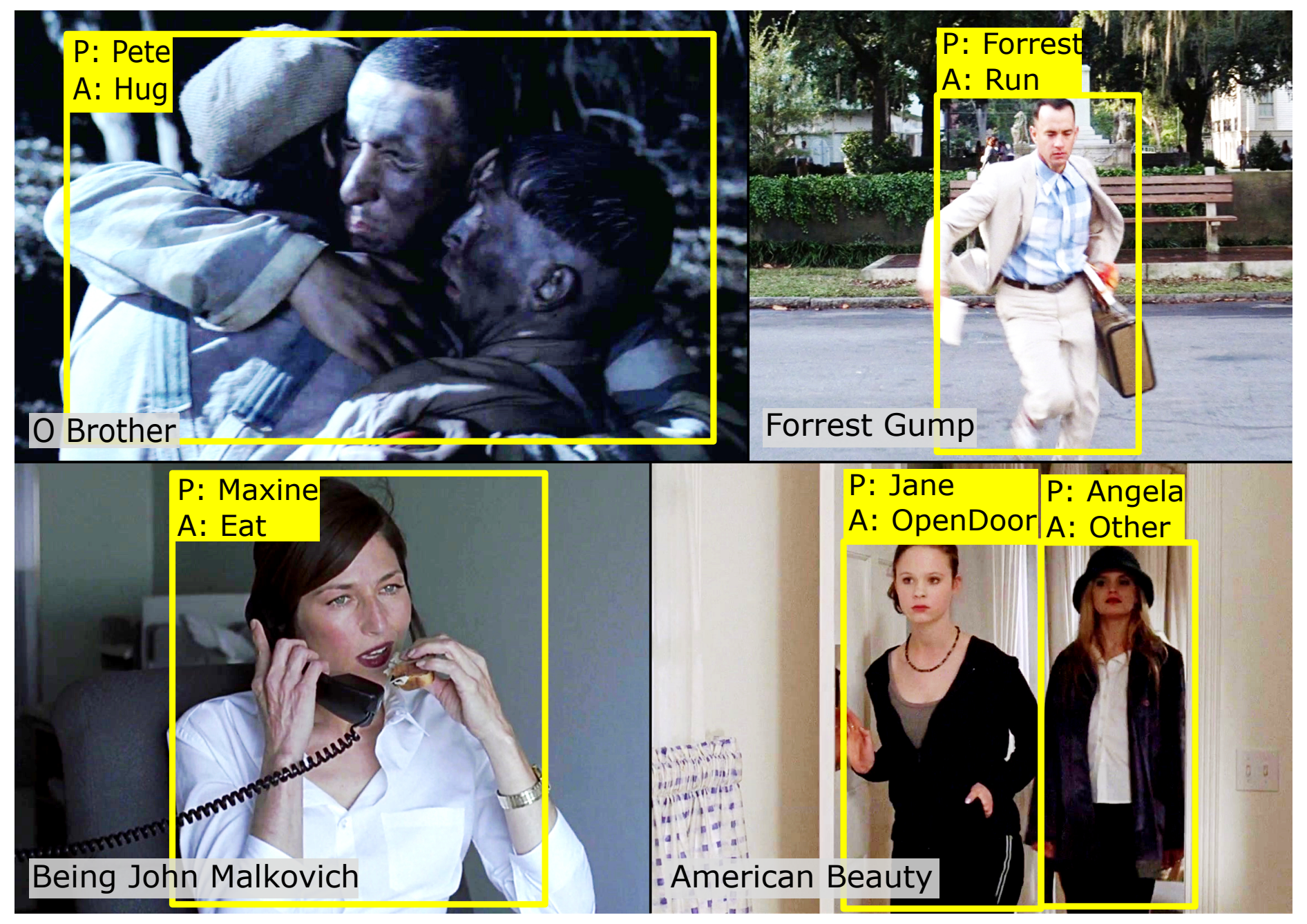
+



Pre-extracted
Person tracks

↓

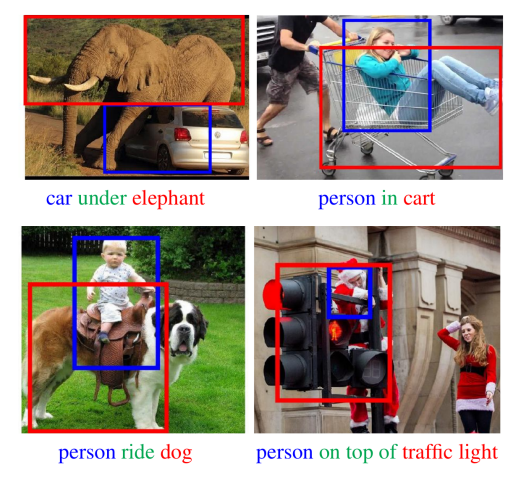
OUTPUT



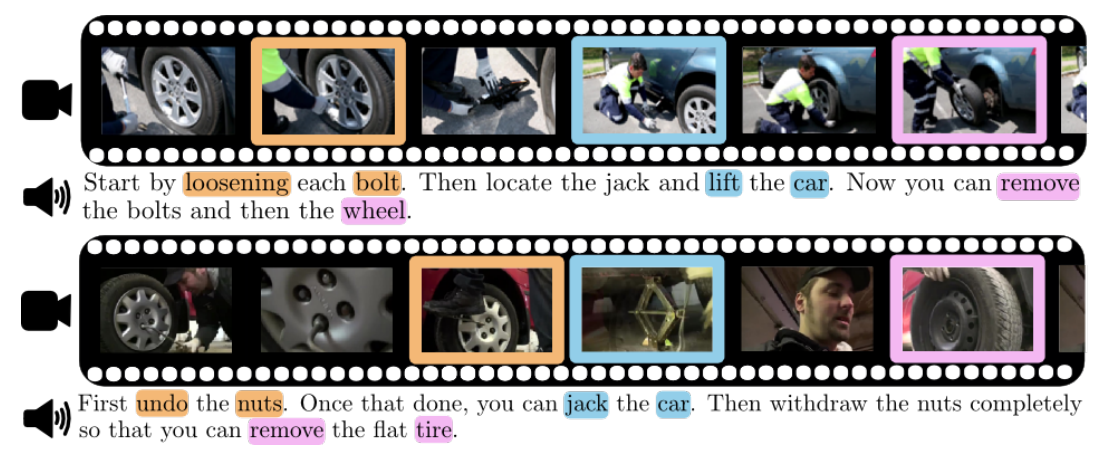
Person + Action recognition

Motivation

Scale-up other weakly-supervised applications:



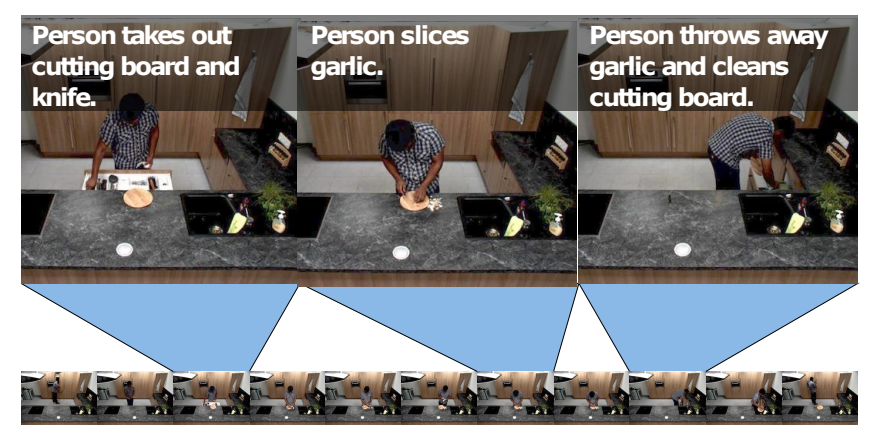
Weakly-supervised learning of visual relations
[Peyre et al. ICCV17]



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.

Unsupervised learning from narrated instructional videos
[Alayrac et al. CVPR16]



Person takes out cutting board and knife.

Person slices garlic.

Person throws away garlic and cleans cutting board.

Weakly-Supervised Alignment of Video with Text
[Bojanowski et al. ICCV15]

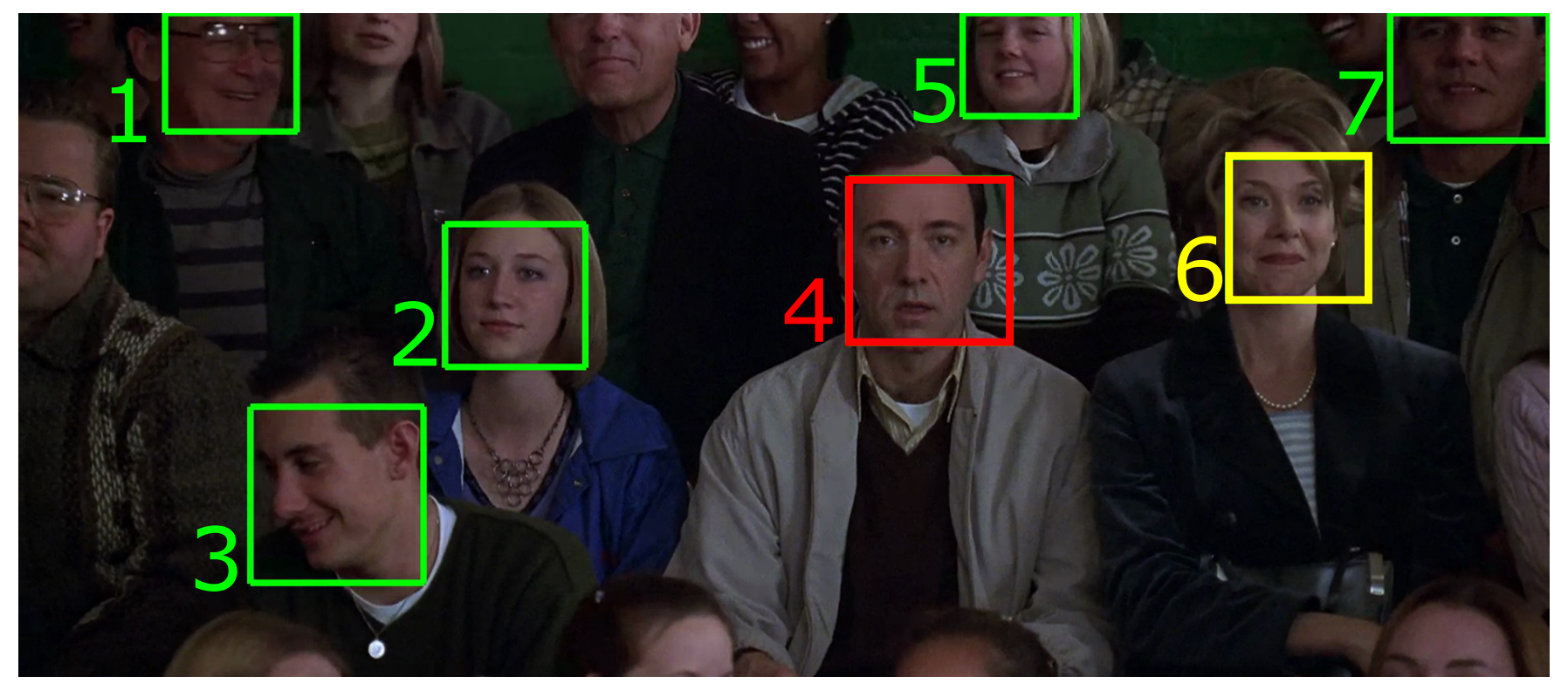
Contributions

- New online optimization algorithm based on Block-Coordinate Frank-Wolfe (BCFW) for scaling-up discriminative clustering
- Improved model of the background class

Discriminative Clustering

$$\min_{Z,W} \frac{1}{N}\|XW - Z\|_F^2 + \lambda\|W\|_F^2$$

- $Z \in \mathbb{R}^{N\times K}$: Assignment matrix (e.g Person name or Action class)
- $X \in \mathbb{R}^{N\times d}$: Person tracks features (e.g VGG-face features for face recognition and Improved Dense Trajectories for Action Recognition)
- $W \in \mathbb{R}^{d\times K}$: Linear model to learn



American Beauty

Scene i Person name parsed from Script: **LESTER**

$$Z^i = \begin{array}{c|ccc|l} & \text{Other} & \text{Lester} & \text{Carolyn} & \\ \hline 1 & 1 & 0 & 0 & \xrightarrow{+} 1 \\ 2 & 1 & 0 & 0 & \xrightarrow{+} 1 \\ 3 & 1 & 0 & 0 & \xrightarrow{+} 1 \\ 4 & 0 & 1 & 0 & \xrightarrow{+} 1 \\ 5 & 1 & 0 & 0 & \xrightarrow{+} 1 \\ 6 & 0 & 0 & 1 & \xrightarrow{+} 1 \\ 7 & 1 & 0 & 0 & \xrightarrow{+} 1 \\ \hline & +\downarrow 5 & +\downarrow 1 & & \end{array}$$

Weak-supervision as Linear Constraints on Z

- **At Least One Constraint**
- **Background Class Constraint**
- **Mutual Exclusion Constraint**

[Bach and Harchaoui, 2007] [Bojanowski et al., 2013]

Large-Scale Optimization

An algorithm based on the Block-Coordinate Frank-Wolfe method for efficient online optimization

$$Z = \boxed{\begin{matrix} Z^1 \\ Z^2 \\ \vdots \\ Z^{N_{movies}} \end{matrix}}$$

- Z is a **block constraint separable** variable
- Exploit the Block-Coordinate Frank-Wolfe algorithm to treat each block in **an online manner**
- Efficient **Time and Space complexity** of block gradient computation via **smart update rules**
- **Convex relaxation** of the problem

Standard FW step
Space complexity: $\mathcal{O}(N^2)$
Time complexity: $\mathcal{O}(N)$

→

Our optimized BCFW step
Space complexity: $\mathcal{O}(N_{block})$
Time complexity: $\mathcal{O}(N_{block})$

Results

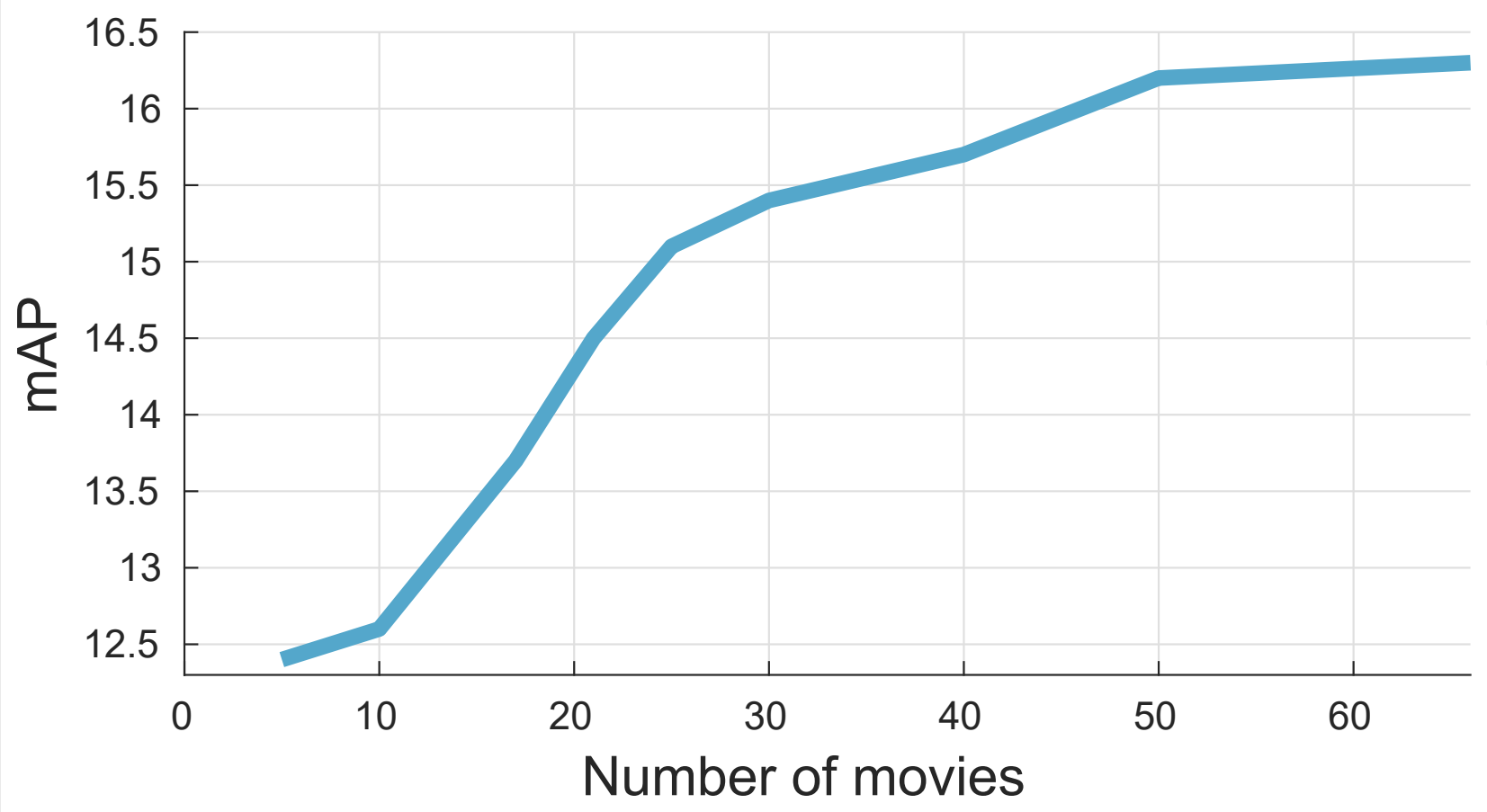
Dataset: 66 feature-length movies together with scripts

Actions: A vocabulary of 14 different actions

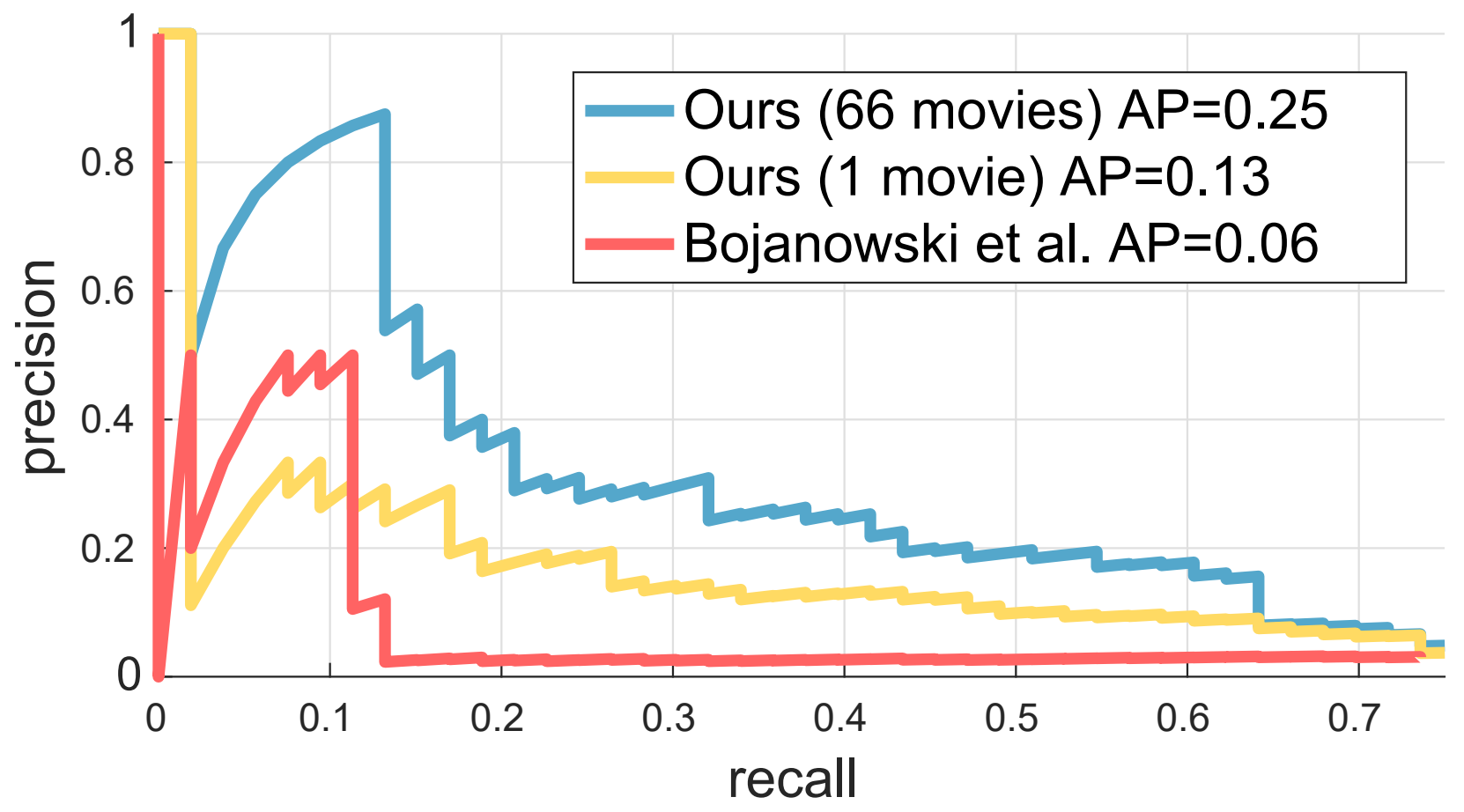
Comparison of different method for Person recognition on *Casablanca*

Method	Acc.	Multi-Class AP	Background AP
Cour et al. [8]	48%	63%	–
Sivic et al. [35]	49%	63%	–
Bojanowski et al. [4]	57%	75%	51%
Parkhi et al. [27]	74%	93%	75%
Our method	**83%**	**94%**	**82%**

Performance when varying the number of training movies



Comparison of different method for the Action Sit Down on *Casablanca*


precision
recall
Ours (66 movies) AP=0.25
Ours (1 movie) AP=0.13
Bojanowski et al. AP=0.06