ERC ALLEGRO workshop
INRIA Grenoble
July 23, 2014

Weakly-supervised learning from videos and scripts

Ivan Laptev

ivan.laptev@inria.fr

WILLOW, INRIA/ENS/CNRS, Paris

Joint work with: Piotr Bojanowski – Rémi Lajugie – Francis Bach – Jean Ponce – Cordelia Schmid – Josef Sivic

Computer vision grand challenge: Dynamic scene understanding

Objects:
cars, glasses, people, etc…

Actions:
drinking, running, door exit, car enter, etc…

constraints

Scene categories:
indoors, outdoors, street scene, etc…

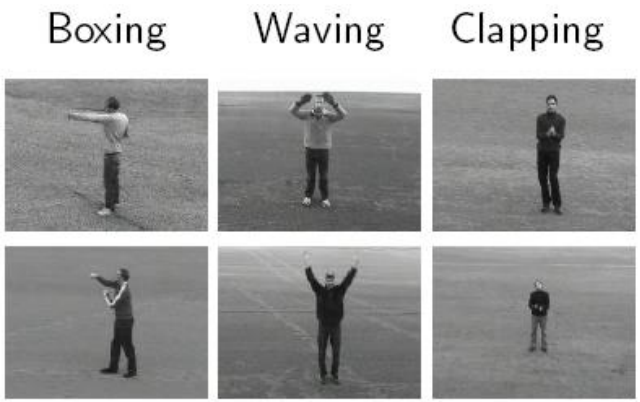
Geometry:
Street, wall, field, stair, etc…

Where to get training data?

- Shoot actions in the lab

  KTH dataset
  Weizman dataset,…

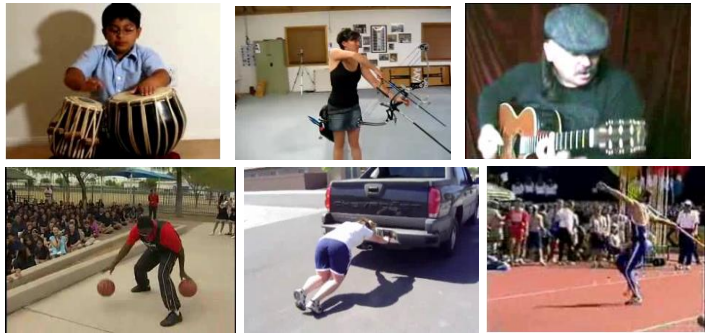
  ⇨ - Limited variability
  - Unrealistic



- Manually annotate existing content

  HMDB, Olympic Sports,
  UCF50, UCF101, …

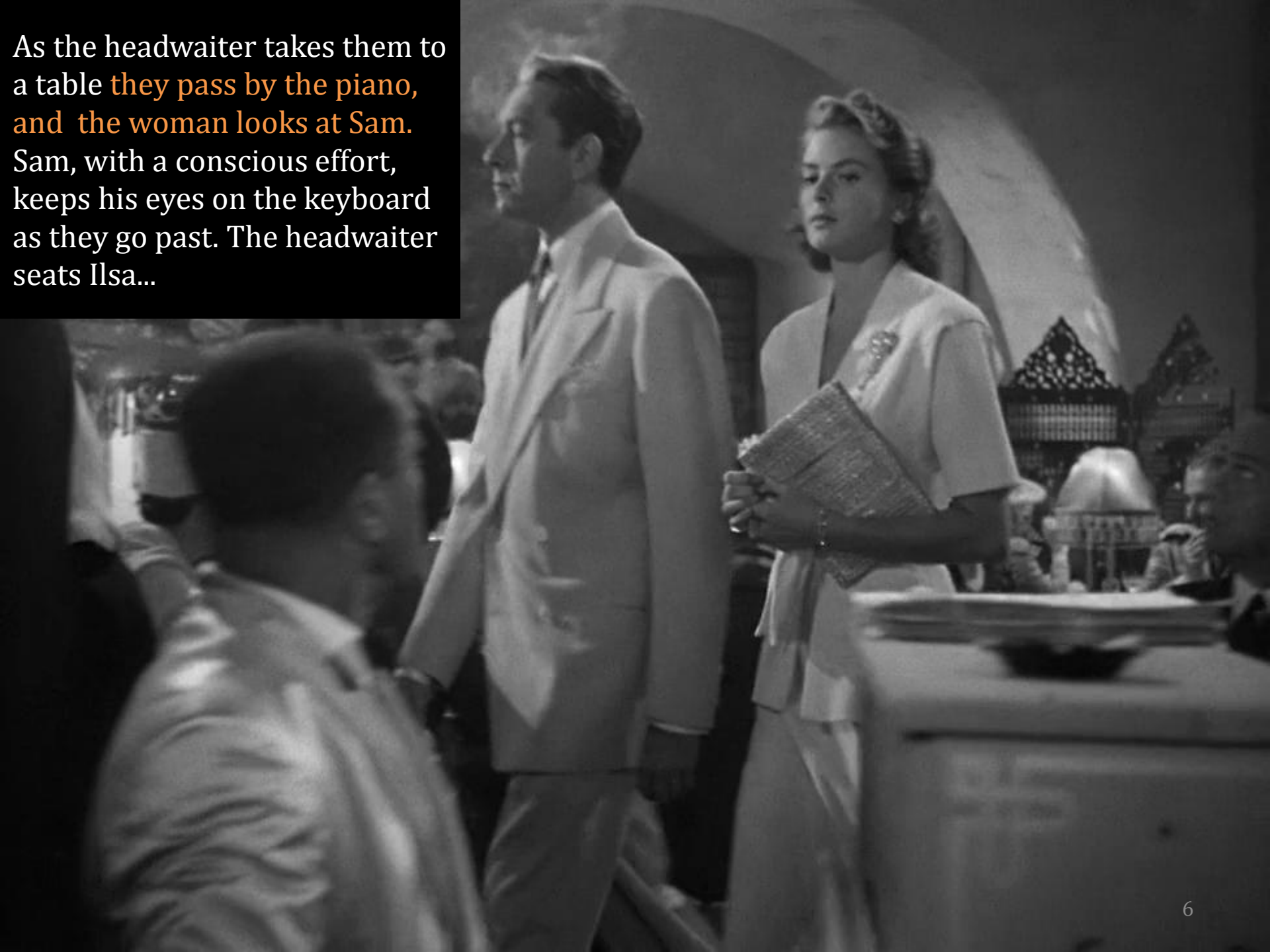
  ⇨ - Very time-consuming

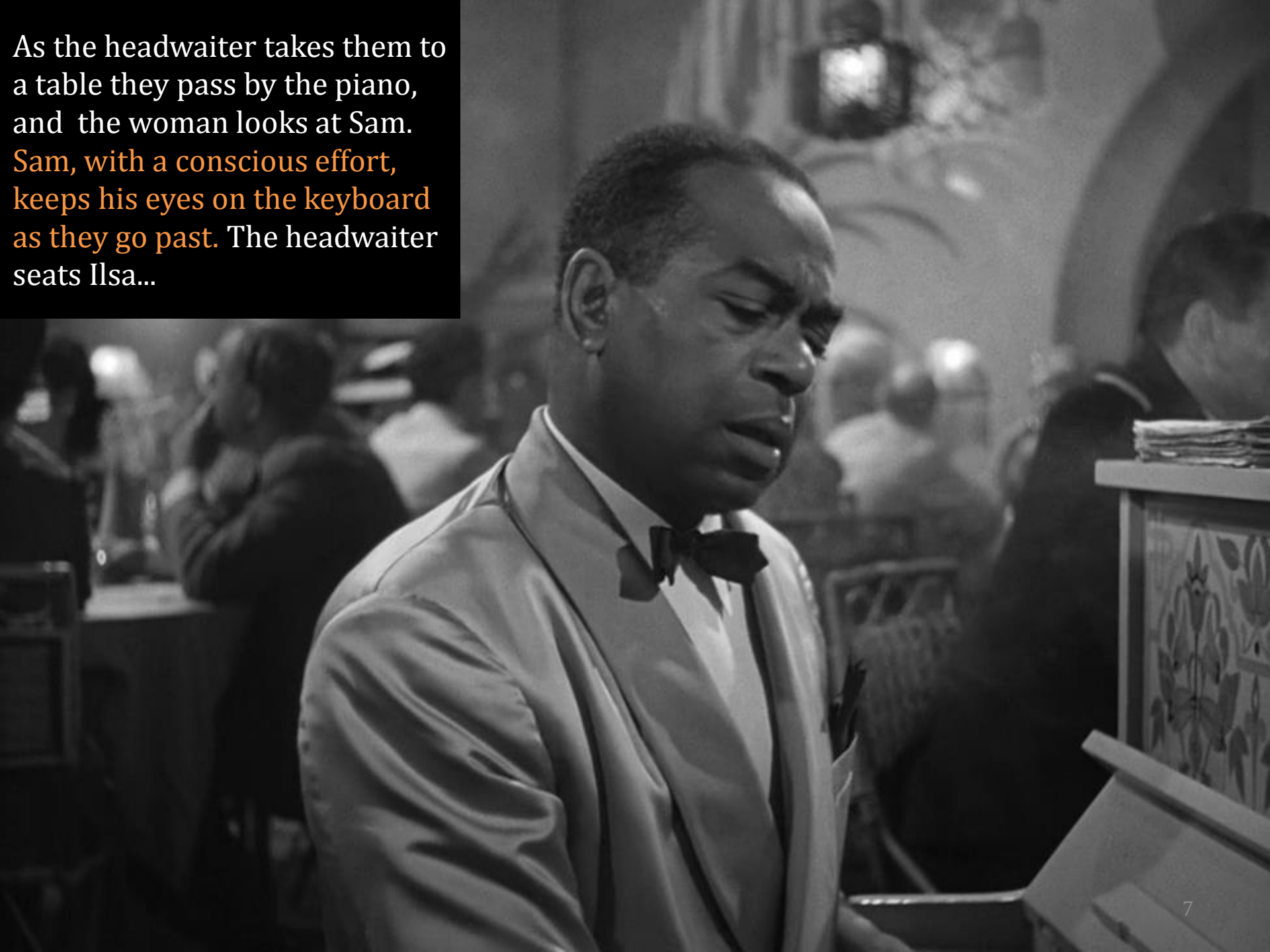


- Use readily-available video scripts
  - Scripts are available for 1000's of hours of movies and TV-series
    www.dailyscript.com, www.movie-page.com, www.weeklyscript.com
  - Scripts describe dynamic and static content of videos


As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

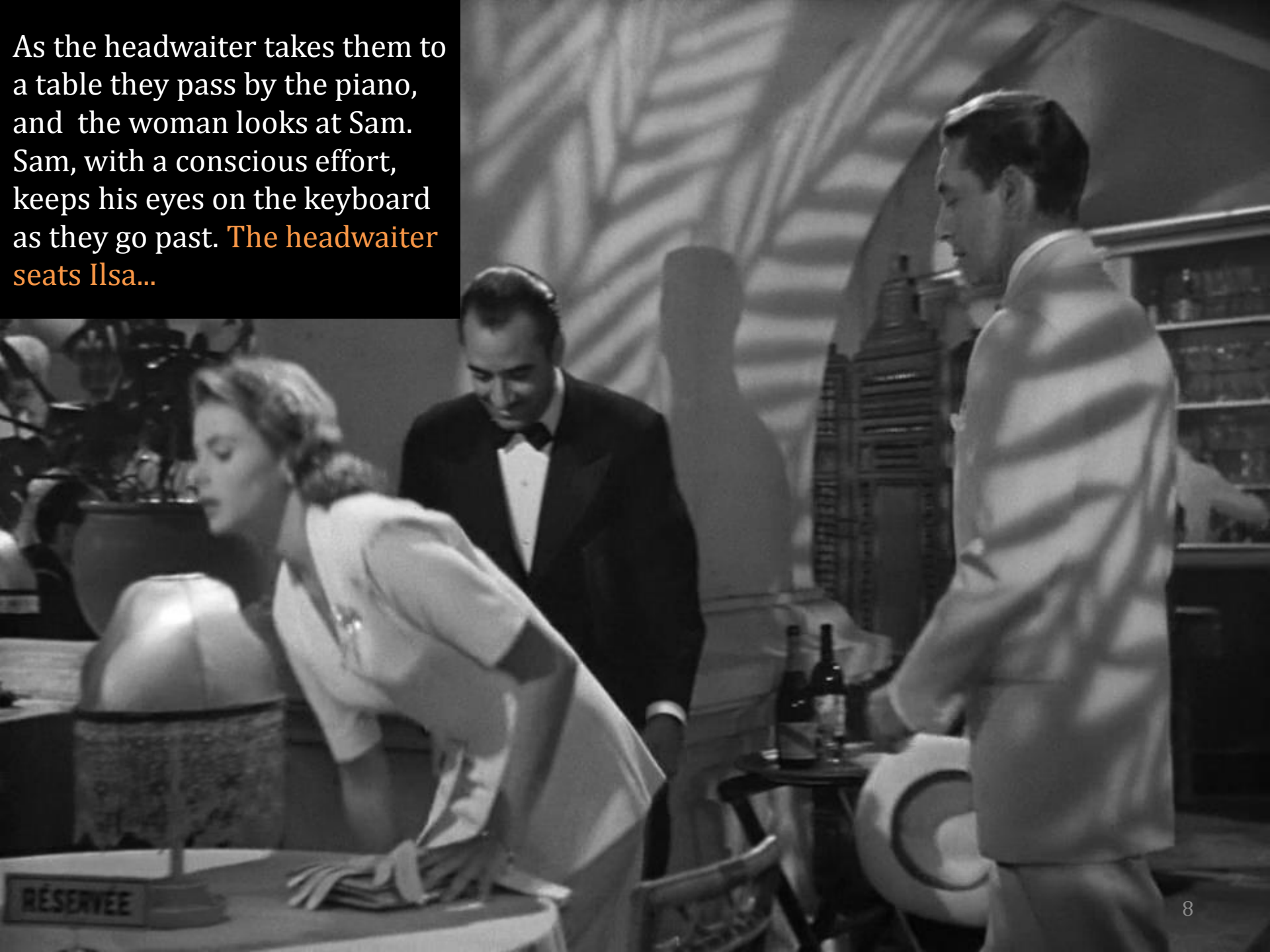
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...





As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...



Scripts as weak supervision

Challenges:

- Imprecise temporal localization
- No explicit spatial localization
- NLP problems, scripts ≠ training labels

"… Will gets out of the Chevrolet. …"
"… Erin exits her new truck…"

vs. Get-out-car

Subtitles

00:24:22 –> 00:24:25
– Yes, Monsieur Laszlo. Right this way.

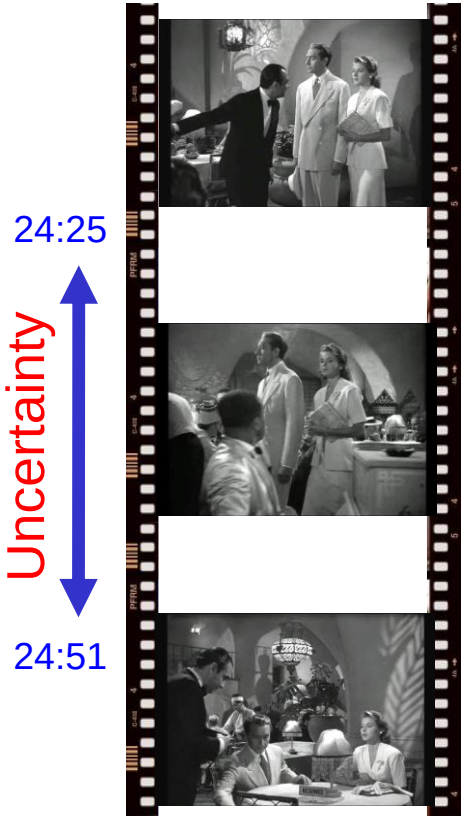
00:24:51 –> 00:24:53
Two Cointreaux, please.

Script

Speech
Monsieur Laszlo. Right this way.

Scene description
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

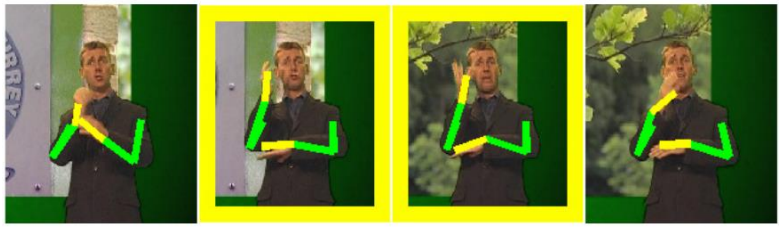
Speech
Two cointreaux, please.



Previous work

Sivic, Everingham, and Zisserman, "Who are you?" -- Learning Person Specific Classifiers from Video, *In CVPR 2009.*

Buehler, Everingham, and Zisserman "Learning sign language by watching TV (using weakly aligned subtitles)", *In CVPR 2009.*



…wanted to know about the history of the ***trees***

Duchenne, Laptev, Sivic, Bach and Ponce, "Automatic Annotation of Human Actions in Video", *In ICCV 2009.*

Joint Learning of Actors and Actions

[Bojanowski et al. ICCV 2013]

Joint Learning of Actors and Actions

[Bojanowski et al. ICCV 2013]

Formulation: Cost function

$$\frac{1}{N}\|Z - \phi(X)w - b\|_F^2 + \lambda_1 \, Tr(w^T \, w)$$

Actor classifier

Actor labels

Actor image features

Rick
Ilsa
Sam

Formulation: Cost function

$$\frac{1}{N}\|Z - \phi(X)w - b\|_F^2 + \lambda_1 \, Tr(w^T \, w)$$

$$\begin{bmatrix} z_{11} & \cdots & z_{1p} & \cdots & z_{1P} \\ \vdots & & \vdots & & \vdots \\ z_{n_1 1} & \cdots & z_{n_1 p} & \cdots & z_{n_1 P} \\ z_{n_2 1} & \cdots & z_{n_2 p} & \cdots & z_{n_2 P} \\ z_{n_3 1} & \cdots & z_{n_3 p} & \cdots & z_{n_3 P} \\ \vdots & & \vdots & & \vdots \\ z_{N1} & \cdots & z_{Np} & \cdots & z_{NP} \end{bmatrix}$$

p = Rick

Weak supervision from scripts:

Person p appears at least once in clip N :

$$\sum_{n \in \mathcal{N}_i} z_{np} \geq 1$$

Formulation: Cost function

$$\frac{1}{N}\|Z - \phi(X)w - b\|_F^2 + \lambda_1 \, Tr(w^T \, w)$$

$$+ \frac{1}{N}\|T - \psi(X)v - c\|_F^2 + \lambda_2 \, Tr(v^T \, v)$$

Weak supervision from scripts:

Action a appears at least once in clip N :

$$\sum_{n \in \mathcal{N}_i} t_{na} \geq 1$$

$$\begin{bmatrix} t_{11} & \cdots & t_{1a} & \cdots & t_{1A} \\ \vdots & & \vdots & & \vdots \\ t_{n_1 1} & \cdots & t_{n_1 a} & \cdots & t_{n_1 A} \\ t_{n_2 1} & \cdots & t_{n_2 a} & \cdots & t_{n_2 A} \\ t_{n_3 1} & \cdots & t_{n_3 a} & \cdots & t_{n_3 A} \\ \vdots & & \vdots & & \vdots \\ t_{N1} & \cdots & t_{Na} & \cdots & t_{NA} \end{bmatrix}$$

a = Walk

Formulation: Cost function

$$\min_{Z,T,w,b,v,c} \quad \frac{1}{N}\|Z - \phi(X)w - b\|_F^2 + \lambda_1 \, Tr(w^T \, w)$$

$$+ \frac{1}{N}\|T - \psi(X)v - c\|_F^2 + \lambda_2 \, Tr(v^T \, v)$$

Weak supervision from scripts:

Person p appears in clip N :

$$\sum_{n \in \mathcal{N}_i} z_{np} \geq 1$$

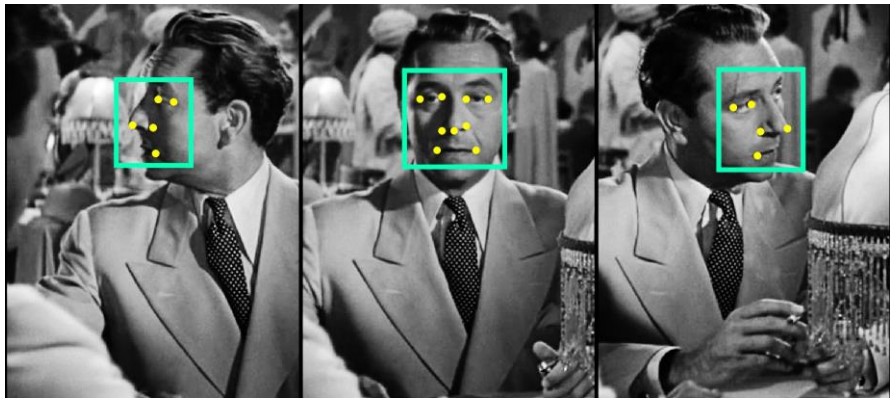
Action a appears in clip N :

$$\sum_{n \in \mathcal{N}_i} t_{na} \geq 1$$

Person p **and** Action a appear in clip N :

$$\sum_{n \in \mathcal{N}_i} z_{np} \, t_{na} \geq 1$$

Image and video features



Face features

$\phi(X)$

- Facial features [Everingham'06]
- HOG descriptor on normalized face image

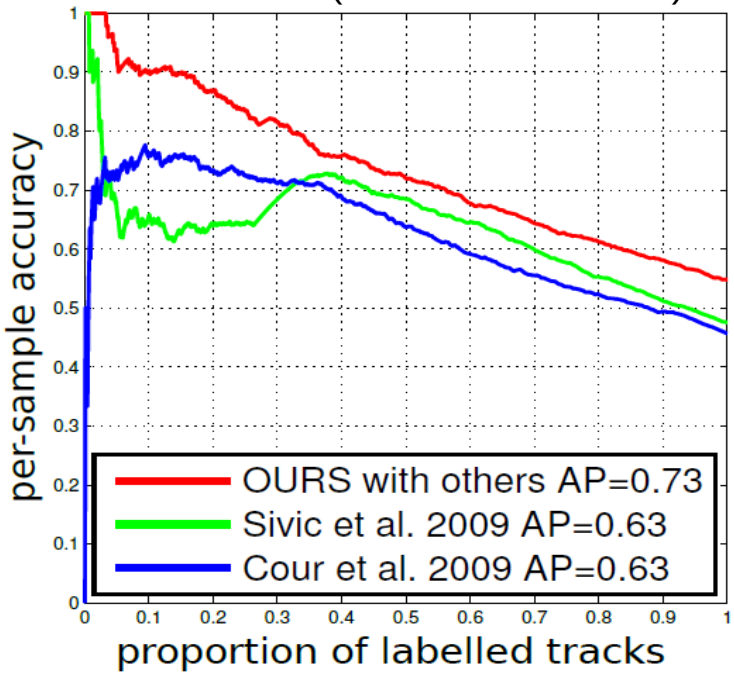
Action features

$\psi(X)$

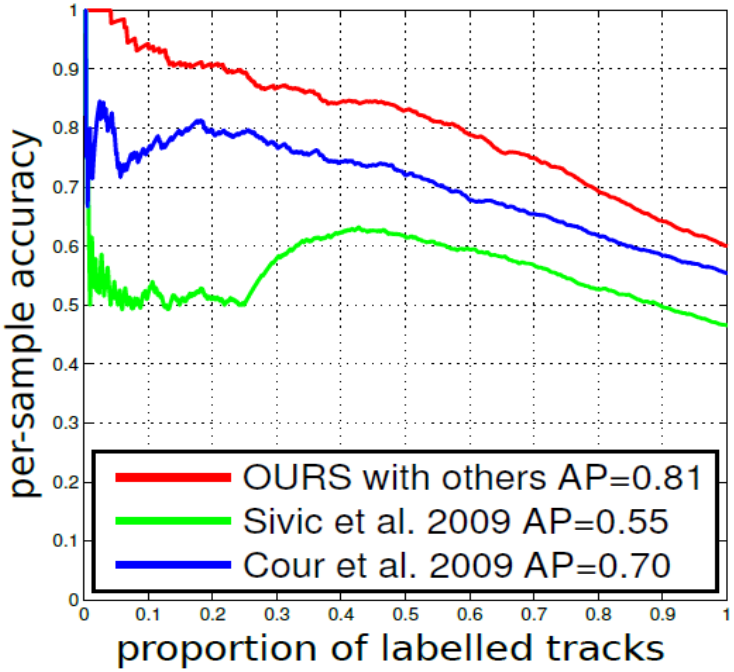
- Dense Trajectory features in person bounding box [Wang et al.,'11]

Results for Person Labelling

Casablanca (17 character names)



American beauty (11 character names)

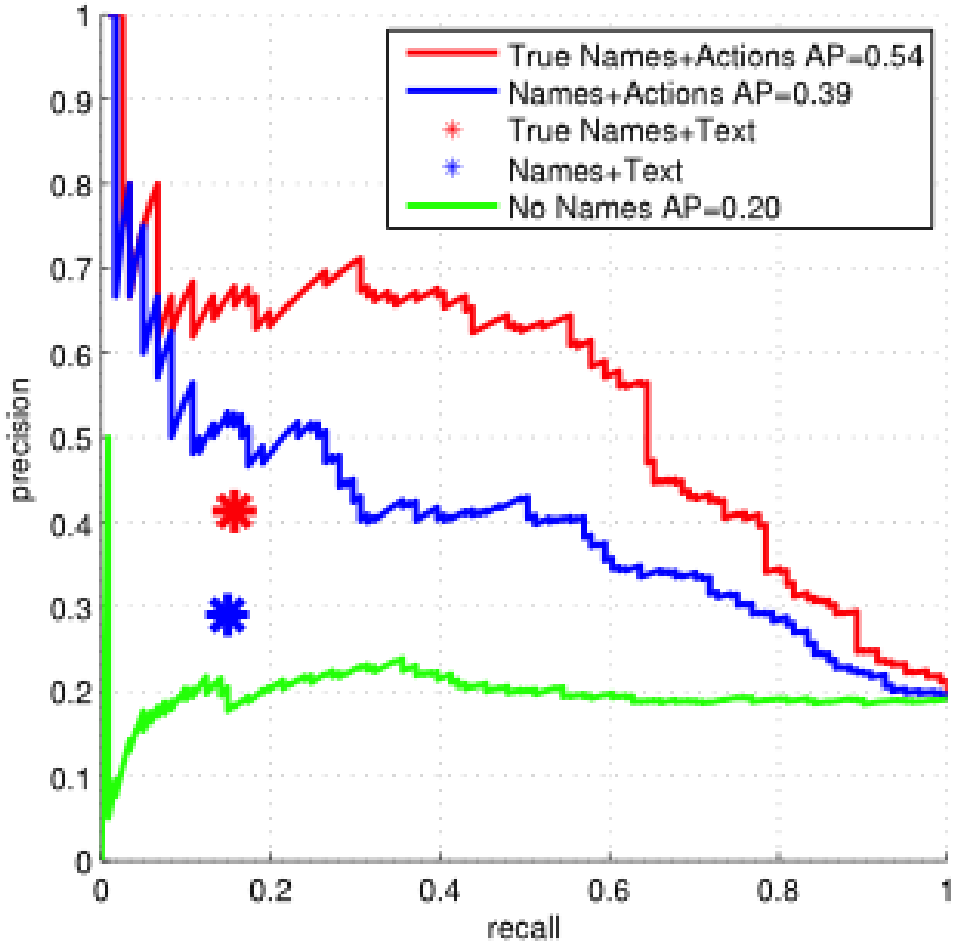


ILSA

RICK

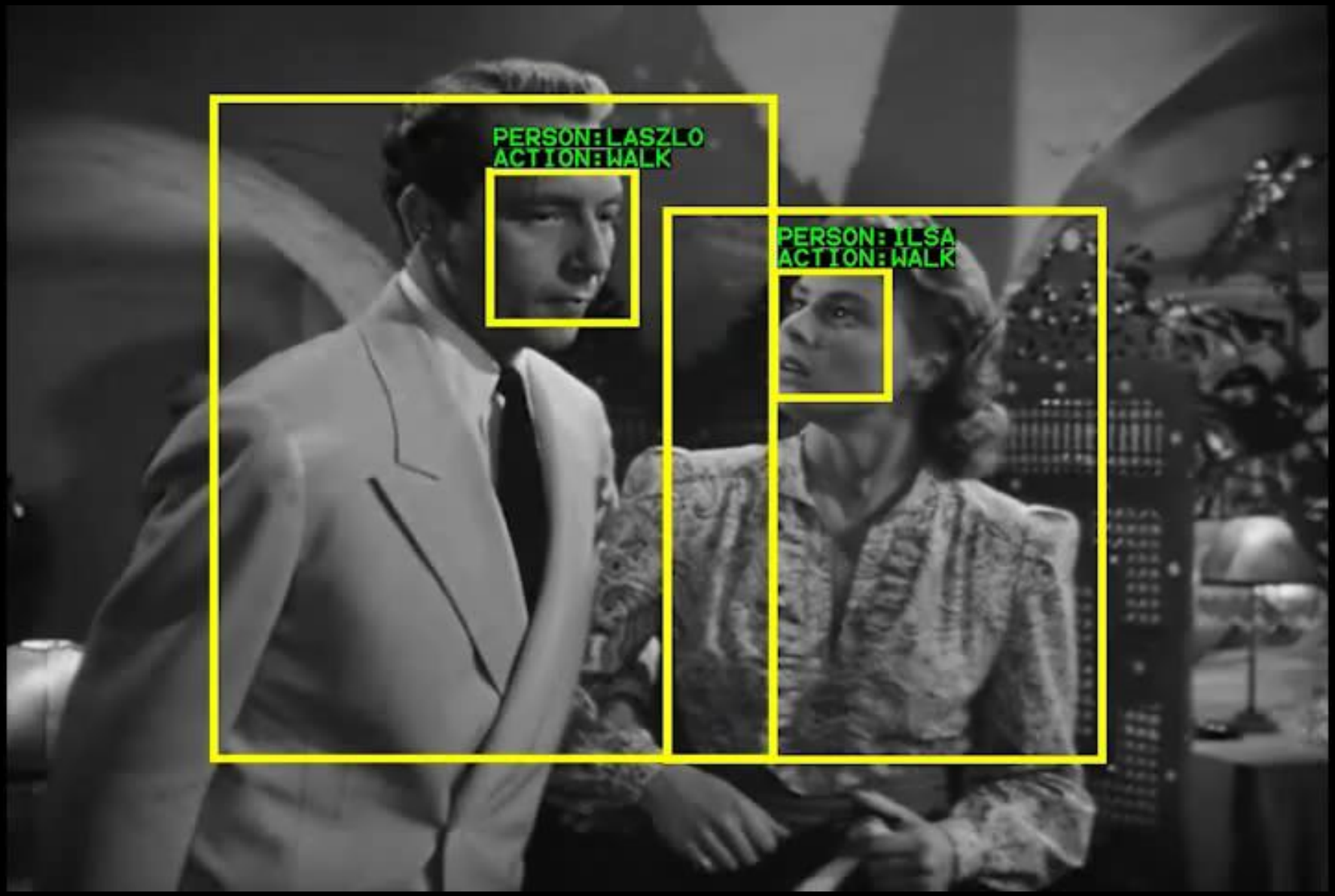
RENAULT

Results for Person + Action Labelling



Casablanca, Walking

Finding Actions and Actors in Movies



[Bojanowski, Bach, Laptev, Ponce, Sivic, Schmid, 2013]

Action Learning with Ordering Constraints

[Bojanowski et al. ECCV 2014]

Action Learning with Ordering Constraints

[Bojanowski et al. ECCV 2014]

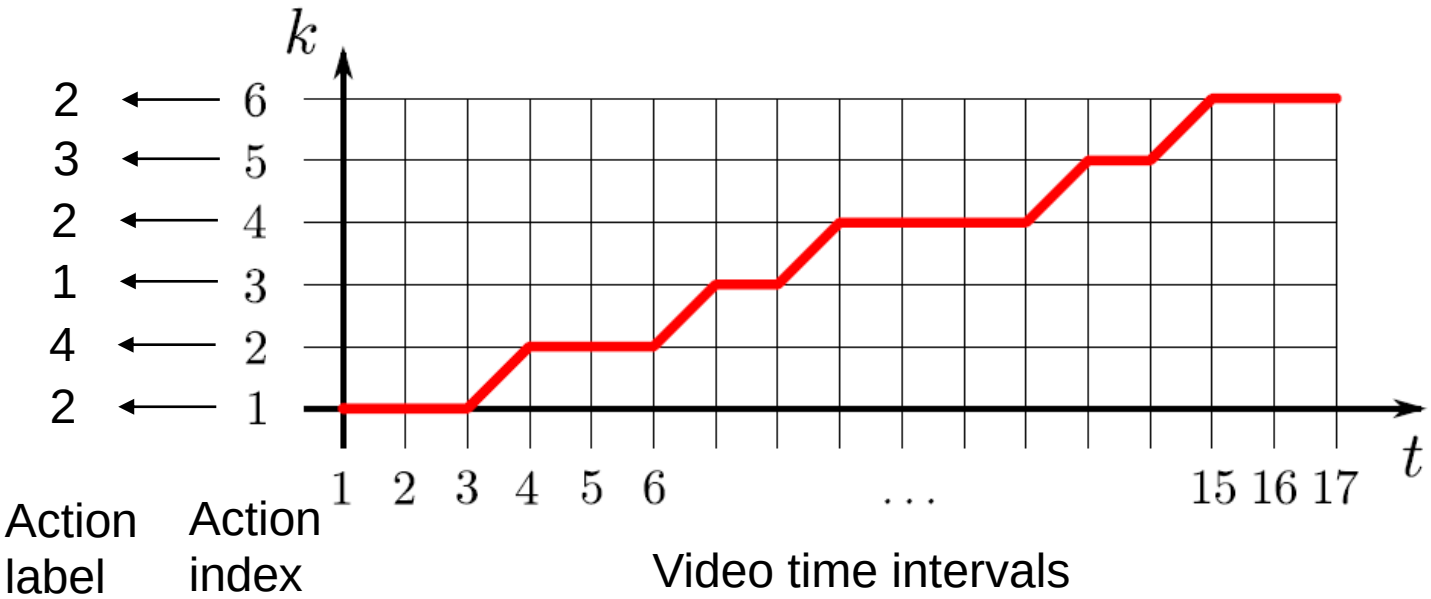
open door → stand up → shake hand

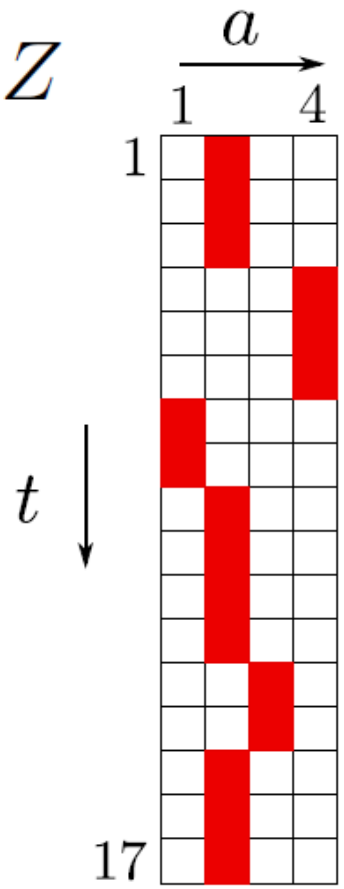
stand up → shake hand → open door

Cost Function

$$\frac{1}{T}\|Z - XW - b\|_F^2 + \frac{\lambda}{2}\|W\|_F^2$$

Weak supervision from ordering constraints on Z:

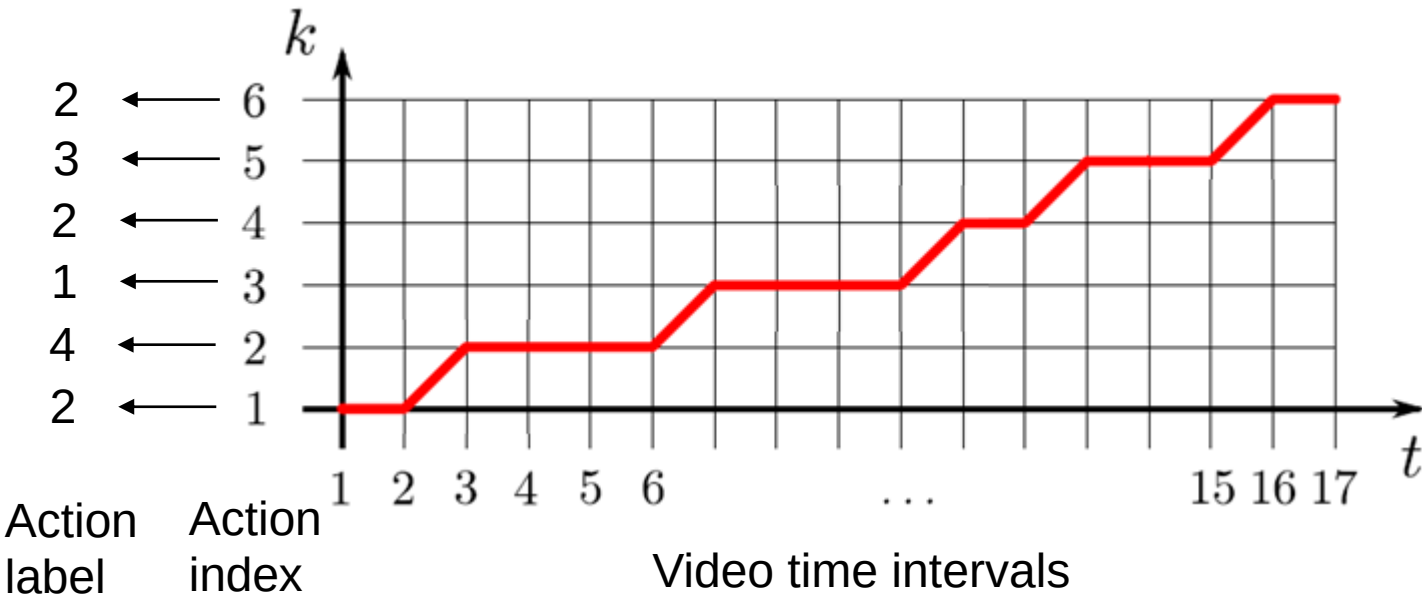


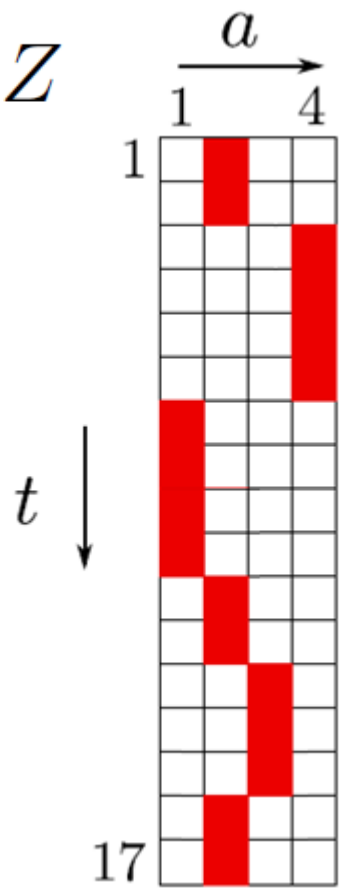


Cost Function

$$\frac{1}{T}\|Z - XW - b\|_F^2 + \frac{\lambda}{2}\|W\|_F^2$$

Weak supervision from ordering constraints on Z:

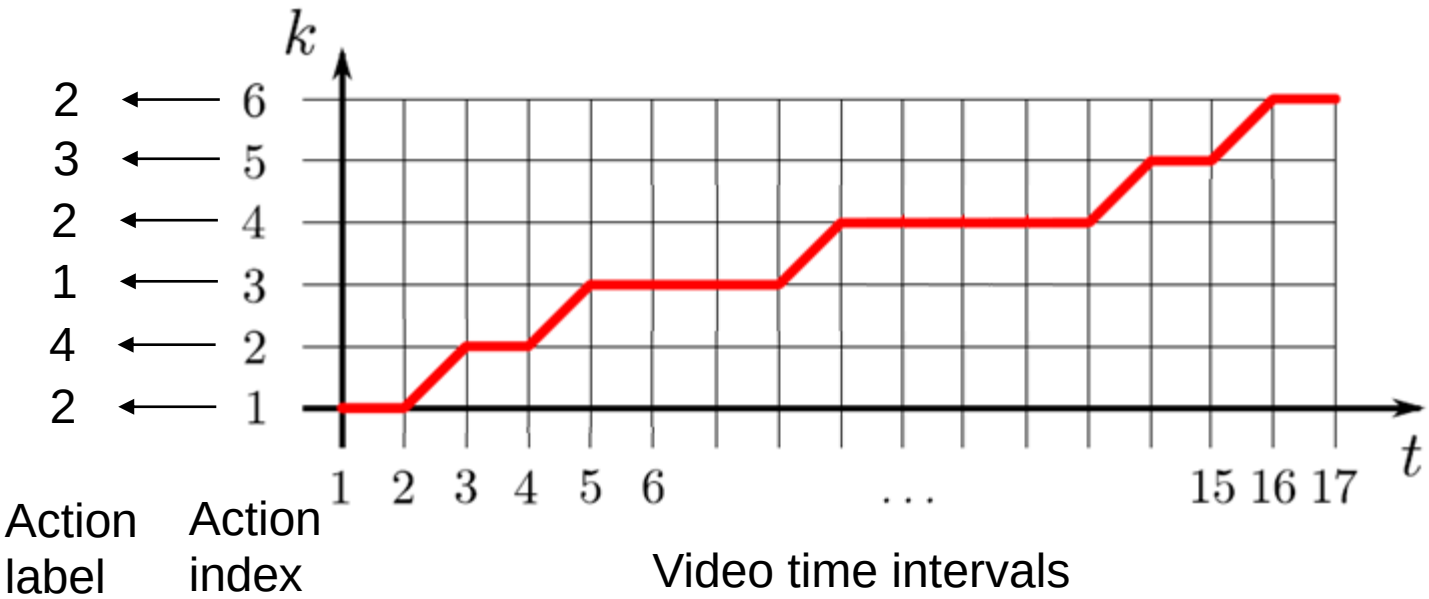


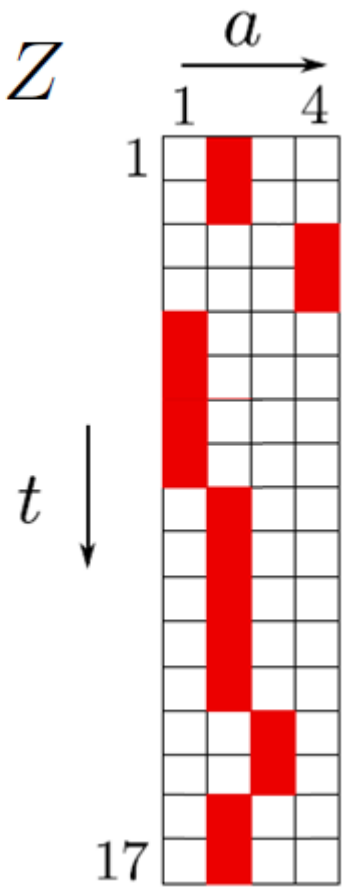


Cost Function

$$\frac{1}{T}\|Z - XW - b\|_F^2 + \frac{\lambda}{2}\|W\|_F^2$$

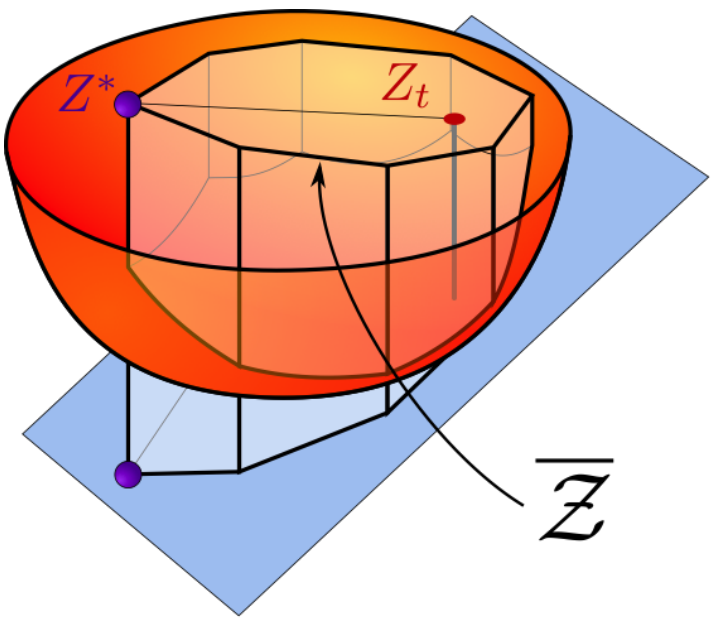
Weak supervision from ordering constraints on Z:

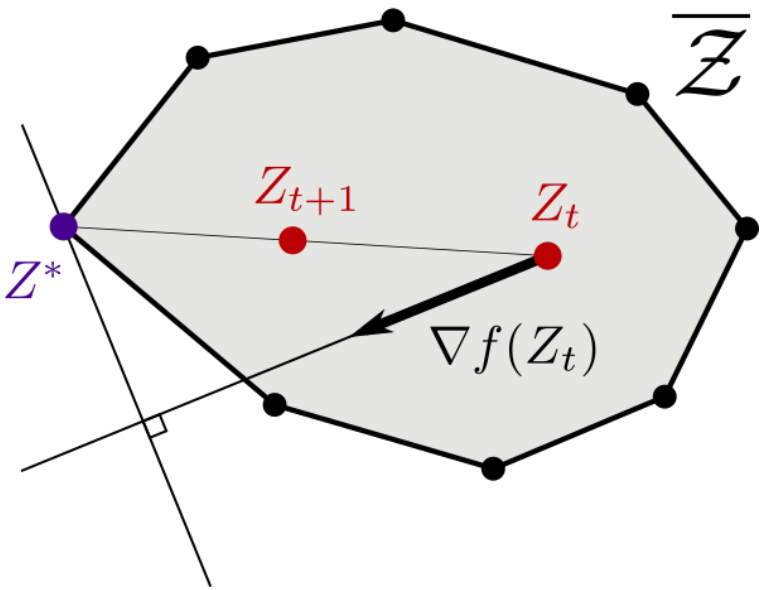




Is the optimization tractable?

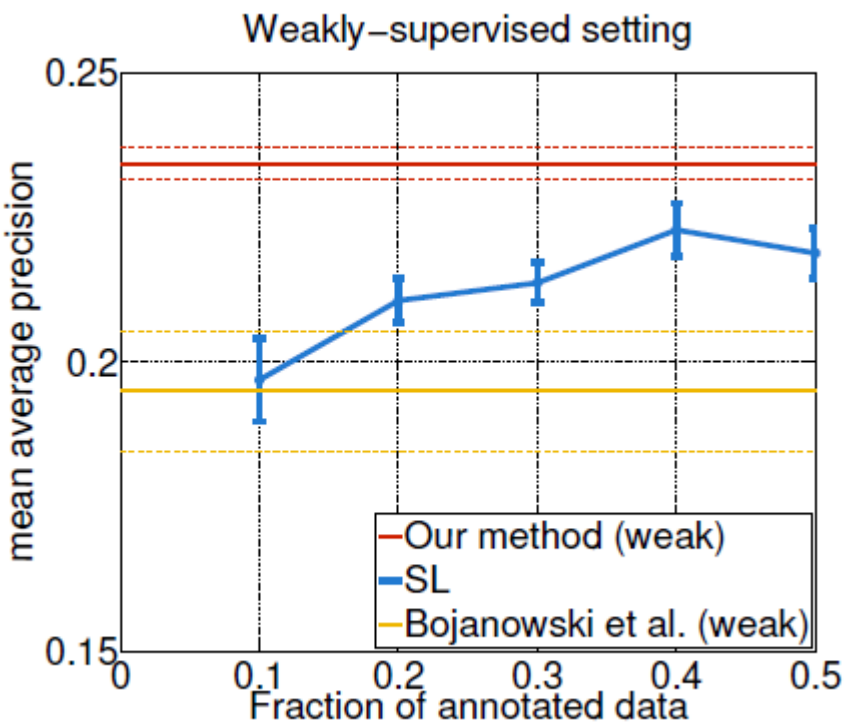
- Path constraints are implicit
- Cannot use off-the-shelf solvers
- Frank-Wolfe optimization algorithm

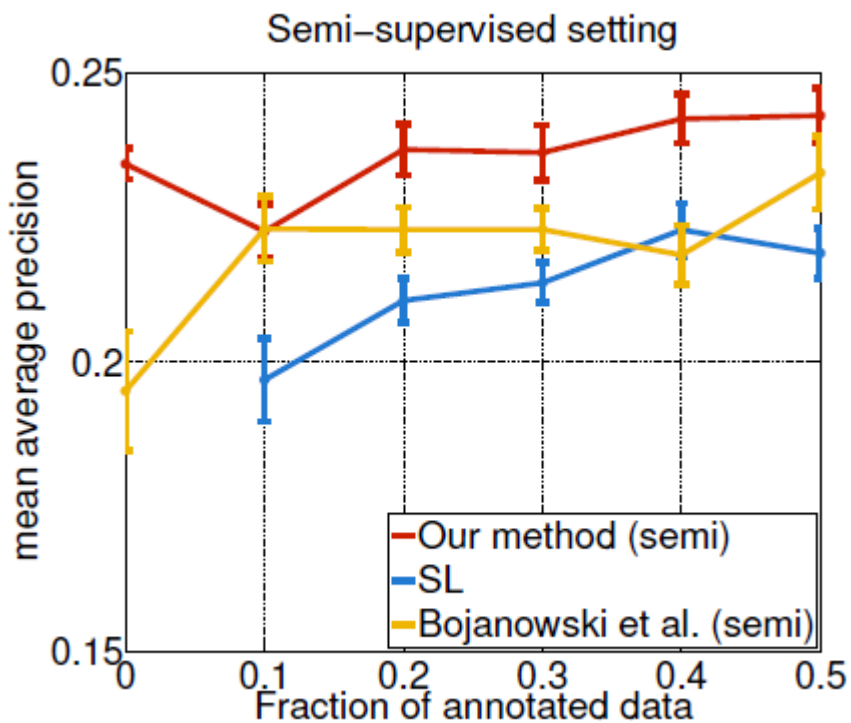




Results

- 937 video clips from 60 Hollywood movies
- 16 action classes
- Each clip is annotated by a sequence of n actions (2≤n≤11)



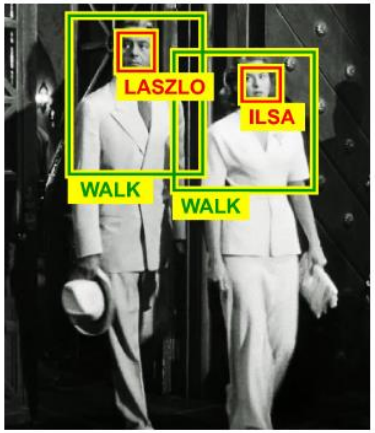


Clip number 0101

Summary

Joint Learning of Actors and Actions

- Reason about individual people.
- Weakly-supervised learning of actions and names.

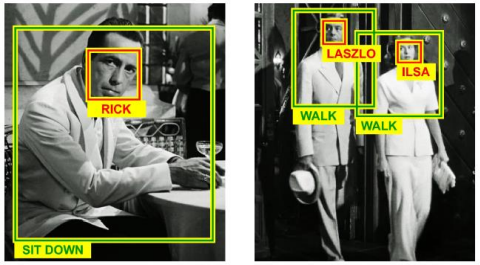


Action learning with ordering constraints

- Reason about action sequences.
- Weakly-supervised learning using time ordering constraints.

Limitations / Future work

Joint Learning of Actors and Actions



- No temporal localization of actions within person tracks.
- Extracting action labels from scripts is a major (NLP+vision?) challenge.
- Finding people in movies is still a big challenge.

Action learning with ordering constraints

- No spatial localization. Want to answer questions:
  - Who is doing what?
  - Who interacts with whom?
- Actions are modeled at short time intervals (15 frames).
- Sequences of action labels are given manually. Want to jointly cluster videos and scripts.