Efficient weakly supervised learning methods in large video collections

Armand Joulin

Stanford University

Linking people in videos with “their” names using coreference resolution

With Vignesh Ramanathan, Percy Liang and Li Fei-Fei

ECCV 2014

Problem setting

- Person naming in TV shows: Assigning name to human tracks

- Problem: No supervision – annotation cost too much

Problem setting

- Instead, we have access to script:

Leonard looks at the robot, while the only engineer in the room fixes it. He is amused.

- Goal: Use this script as a source of **weak supervision**

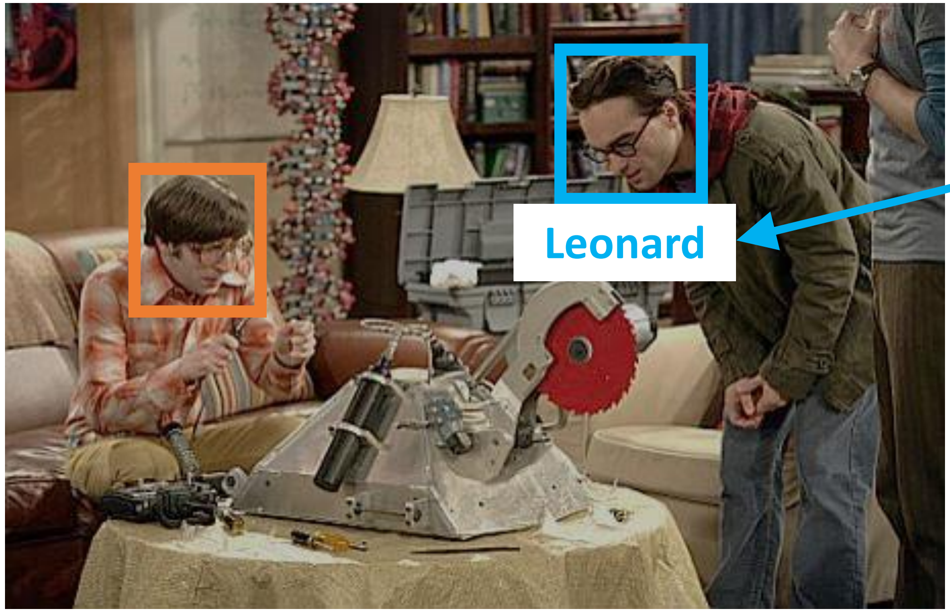
Previous work

- In Bojanowski *et al.* (2013), they extract names from the script:

Leonard looks at the robot, while the only engineer in the room fixes it. He is amused.

Previous work

- In Bojanowski *et al.* (2013), they extract names from the script:



Leonard looks at the robot, while the only engineer in the room fixes it. He is amused.

- Problems:
  - people not always explicitly mentioned
  - Script is a temporal sequence

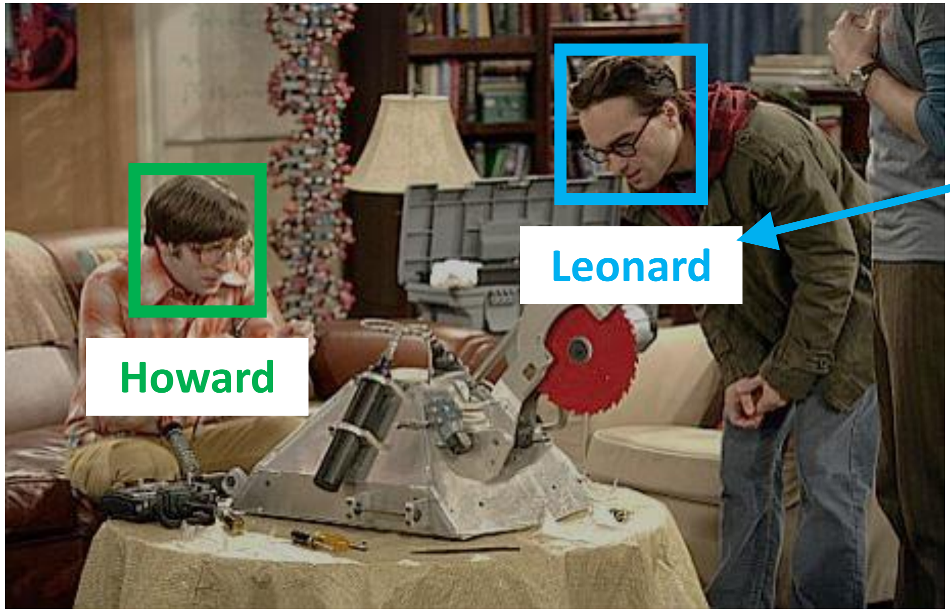
Can we do better?

- Let’s consider all **mentions** of humans in the script:

Leonard looks at the robot, while the only **engineer** in the room fixes it. **He** is amused.

Can we do better?

- Let's consider all **mentions** of humans in the script:



Leonard looks at the robot, while the only engineer in the room fixes it. He is amused.

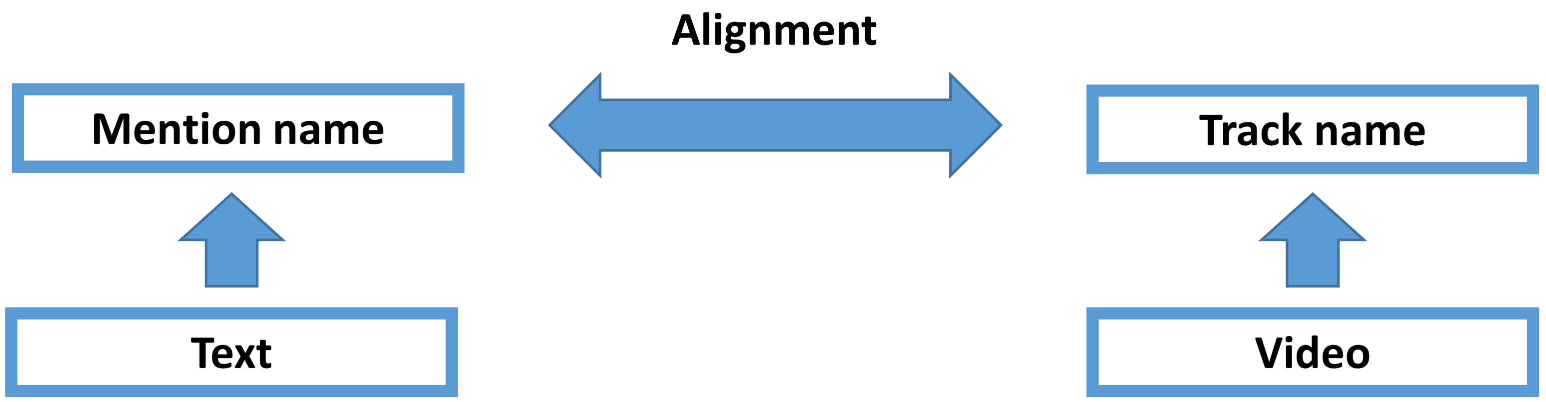
?

- Challenge: Requires to resolve identity of all mentions, i.e., Coreference resolution

Our approach

- We propose a model which **jointly** tackle two problems:
  - A **vision** problem: Track naming
  - A **NLP** problem: Coreference resolution
- We show improvement on both tasks

Our approach



- Difficulty: Text and video are not directly comparable
- Instead:
  - Infer name associated with mention (coreference)
  - Infer name associated with track (track naming)
  - Align them following temporal ordering (alignment)

What is this coreference resolution?

- **Coreference resolution**: Resolve the identity of **ambiguous** mentions (e.g., “he”, “engineer”) by finding **indirectly** a **unambiguous** mention appearing **previously** in the text
- For example:

Roland arrives. He looks foreign. Ian waits as the foreigner rides up

What is this coreference resolution?

- **Coreference resolution**: Resolve the identity of **ambiguous** mentions (e.g., “he”, “engineer”) by finding **indirectly** a **unambiguous** mention appearing **previously** in the text
- For example:

Roland arrives. **He** looks foreign. **Ian** waits as the **foreigner** rides up

What is this coreference resolution?

- **Coreference resolution**: Resolve the identity of **ambiguous** mentions (e.g., “he”, “engineer”) by finding **indirectly** a **unambiguous** mention appearing **previously** in the text
- For example:

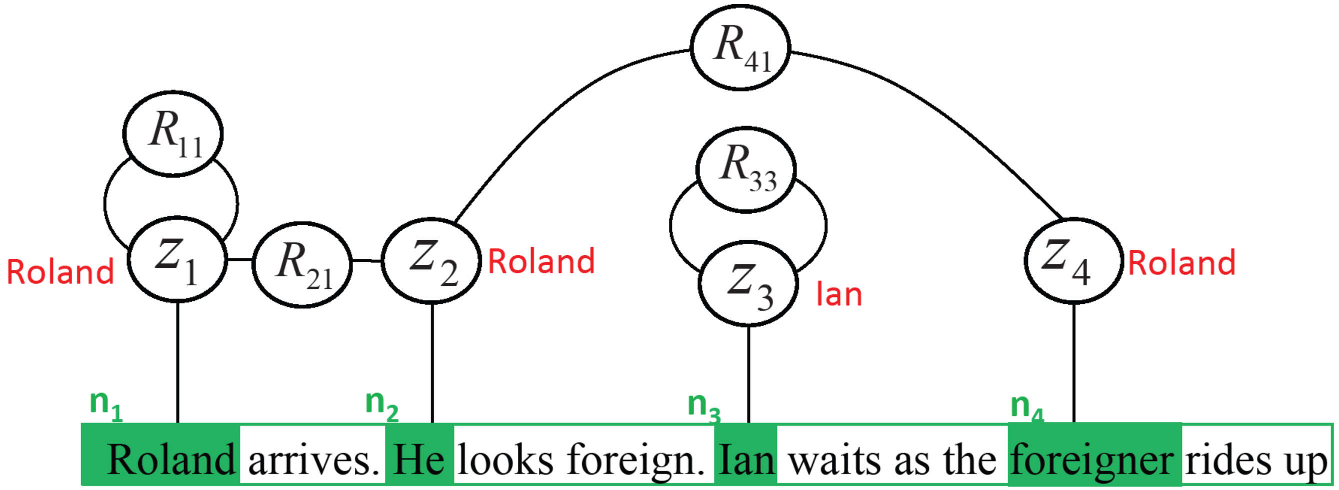
Roland arrives. **He** looks foreign. **Ian** waits as the **foreigner** rides up

What is this coreference resolution?

- **Coreference resolution**: Resolve the identity of **ambiguous** mentions (e.g., “he”, “engineer”) by finding **indirectly** a **unambiguous** mention appearing **previously** in the text
- For example:

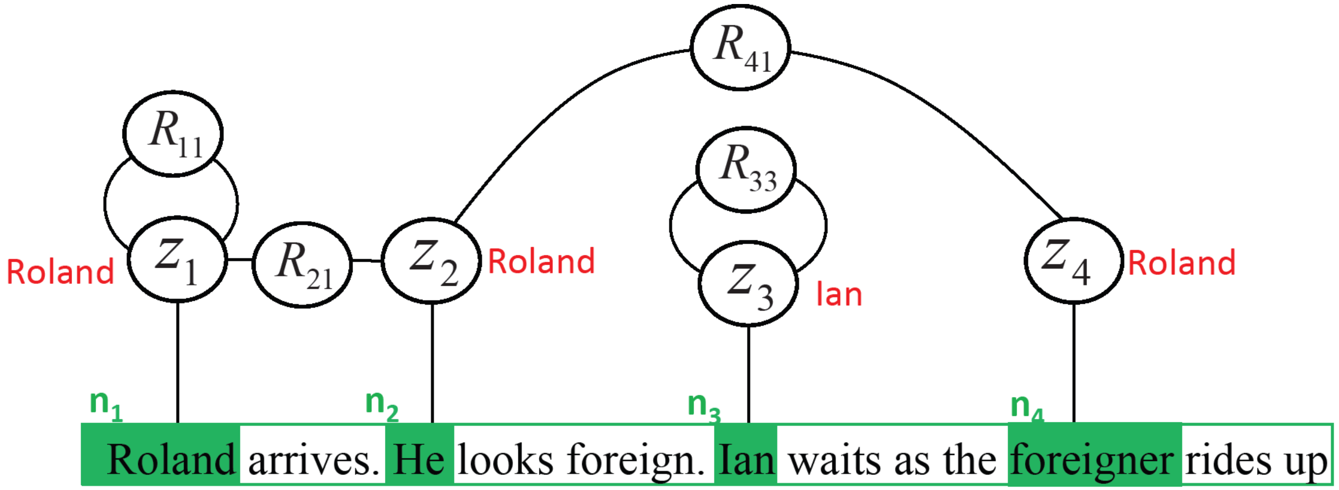
Roland arrives. He looks foreign. Ian waits as the foreigner rides up

Formulation for coreferencing



- Each **pair of mentions** is associated with:
  - A feature x
  - A link variable R in {0,1}
- Each **mention** is associated with:
  - A name variable Z

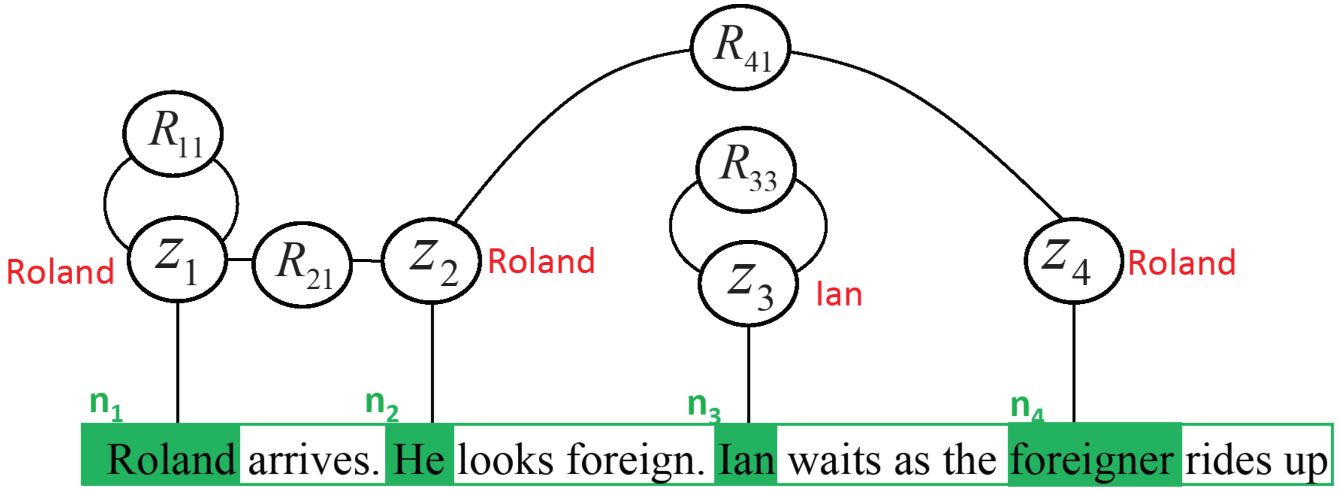
Formulation of coreferencing



- We learn a discriminative model over the mention relation:

$$\underset{R \in \mathbb{P}_{NN},\ w_c,\ b_c}{\text{minimize}} \sum_{n=1}^{N} \sum_{m \leq n} (R_{nm} - x_c^{nm} \cdot w_c - b_c)^2 + \lambda_c \|w_c\|_2^2.$$

Formulation of coreferencing

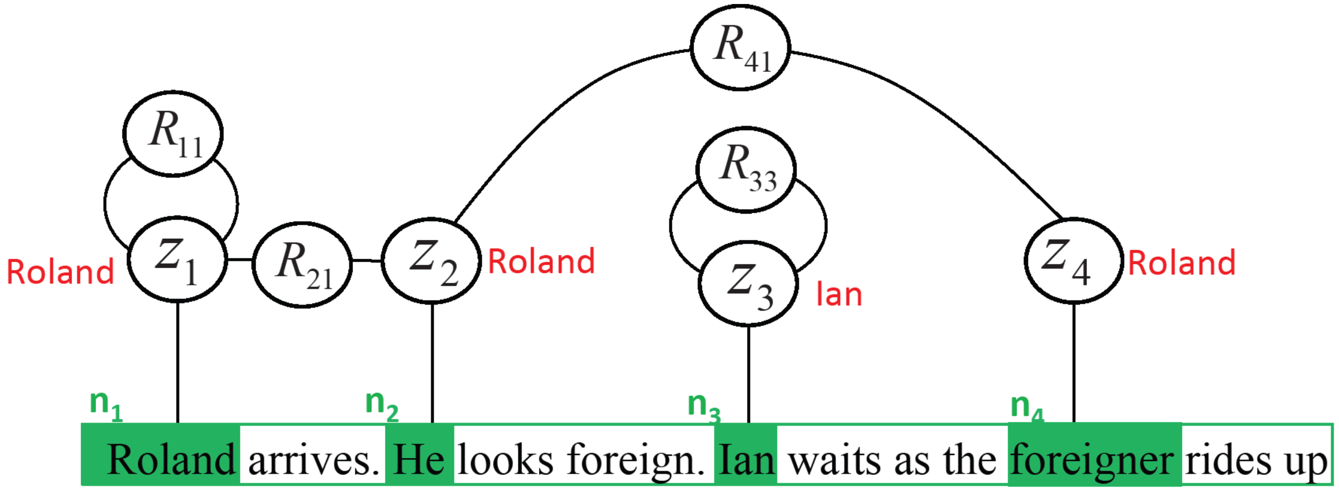


- This problem is in closed form in w and b :

$$\underset{R \in \mathbb{P}_{NN},\ Z \in \mathbb{P}_{NP}}{\text{minimize}} \qquad \text{vec}(R)^T A_C \text{vec}(R)$$

- Where A is an sdp matrix (see Bach and Harchaoui, 2008)

Formulation of coreferencing



- Adding the constraints of coreferencing we have:

$$
\begin{array}{lll}
\underset{R \in \mathbb{P}_{NN},\ Z \in \mathbb{P}_{NP}}{\text{minimize}} & \mathrm{vec}(R)^T A_C \mathrm{vec}(R) & \\
\text{subject to} & \forall\, n \le N,\ \sum_{m \le n} R_{nm} = 1, & \text{(antecedent)} \\
& \forall\, m \le n,\ \|Z_n - Z_m\|_\infty \le 1 - R_{nm} & \text{(connection).}
\end{array}
$$

Formulation for track naming

- x : feature associated with a track
- y : name assignment of a track
- We use the same formulation as in our coreference resolution model.

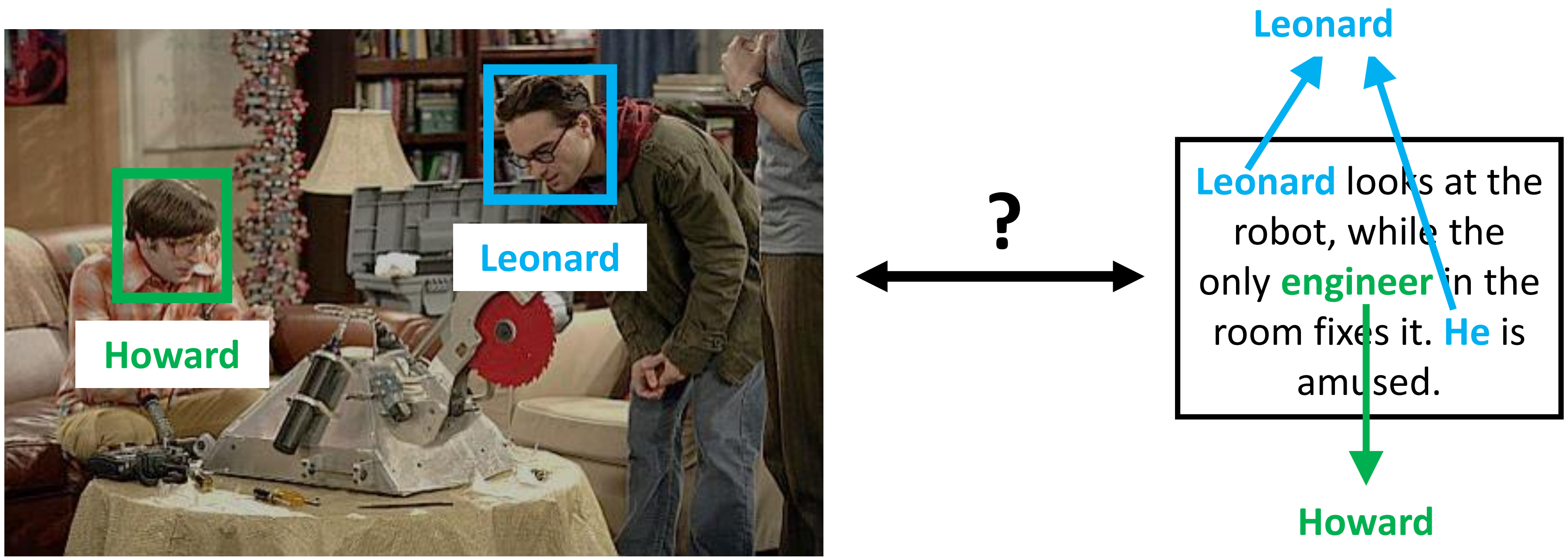
Formulation for track naming

- This leads to a similar IQP (similar to Bojanowski *et al.,* 2013):

$$\begin{aligned}
\underset{Y \in \mathbb{P}_{NP}}{\text{minimize}} \quad & \operatorname{tr}\left(Y^T A_T Y\right) \\
\text{subject to} \quad & \forall\, d \in \mathcal{D},\ \forall\, p \in \mathcal{P}_d \ \sum_{t \in \mathcal{T}_d} Y_{tp} \geq 1, \text{ (dialogue alignment)} \\
& \forall s \in \mathcal{S},\ p \notin \mathcal{P}_s,\ \sum_{t \in \mathcal{T}_s} Y_{tp} = 0. \quad \text{(scene alignment)}
\end{aligned}$$

Where Y is the matrix of all name assignment variables.

Mapping between tracks and mentions



- To ensure a flow of information between text and video, we need to **align the tracks to the mentions**
- We align tracks and mentions based on their name and temporal ordering

Mapping between tracks and mentions

- We align the track name variable Y to the mention one, Z:

$$\begin{aligned} \underset{M \in \{0,1\}^{T \times N}}{\text{minimize}} \quad & \|M^T Y - Z\|_F^2 \\ \text{subject to} \quad & \forall\, e \in \mathcal{E},\ n \in \mathcal{N}_e,\ \sum_{t \in \mathcal{T}_e} M_{tn} = 1 \qquad \text{(mention mapping)} \\ & \forall\, n < N,\ t \leq T\ ,\ \sum_{s=1}^{t} M_{sn} \geq \sum_{s=1}^{t} M_{sn+1}\ \text{(temporal ordering)}. \end{aligned}$$

where M is the alignment variable

- Constraints on Y and Z => $\|M^T Y - Z\|_F^2 = -2\mathrm{tr}(M^T Y Z)$ + Cste

Overall model

- Adding the coreference, track naming and alignment terms, we have:

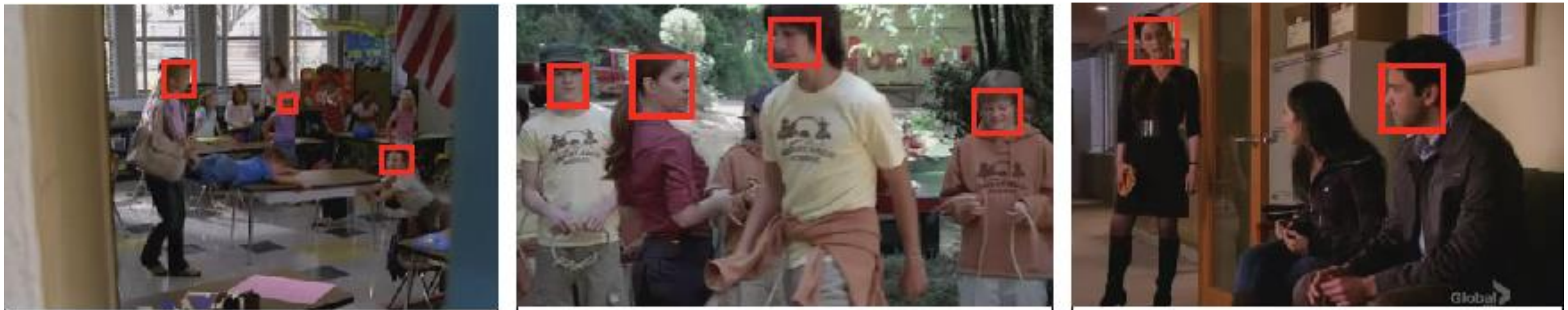
$$\gamma_f \operatorname{tr}\left(Y^T A_T Y\right) + \gamma_s \operatorname{tr}\left(\operatorname{vec}(R)^T A_C \operatorname{vec}(R)\right) - 2\operatorname{tr}(M^T Y Z)$$

Where the parameters are fixed on a validation set.

- We relax it by replacing {0,1} by [0,1]
- We alternate minimization in Y, (Z,R) and M
- The minimization in M can be done by dynamic programing.

Results

- We introduce a databases of 19 TV episodes (+scripts) taken randomly form 10 different TV series
- We run a standard face detector and tracker.
- We only consider human mention which are subject of a verb



We reveal Lynette holding Porter by his feet, while he clings to Preston's desk.

Missy points to the larger kid. The big kid walks off. Other kids jeer.

Cary eyes the siblings, as Alicia looks across the bullpen

Results on track naming

Set	Development				Test					
Episode id	E1	E2	E3	*mAP*	E15	E16	E17	E18	E19	*mAP*
Rand. Chance	0.266	0.254	0.251	0.257	0.177	0.217	0.294	0.214	0.247	0.229
Cour [6]	0.380	0.333	0.393	0.369	0.330	0.327	0.342	0.306	0.337	0.328
Boj. [9]	0.353	0.434	0.426	0.404	0.285	0.429	0.378	0.383	0.454	0.385
Our (flat)	0.512	0.560	0.521	0.531	0.340	0.474	0.503	0.399	**0.384**	0.420
Our+flat cor.	0.497	0.572	0.501	0.523	0.388	0.470	0.512	0.424	**0.401**	0.431
Our+uni.	0.497	0.552	0.561	0.537	0.345	**0.488**	0.516	0.410	0.388	0.429
Our+rand.	0.499	0.497	0.532	0.509	0.344	0.480	0.511	0.404	0.367	0.428
Our (full)	**0.551**	**0.641**	**0.641**	**0.611**	**0.402**	0.483	**0.576**	**0.465**	0.382	**0.461**

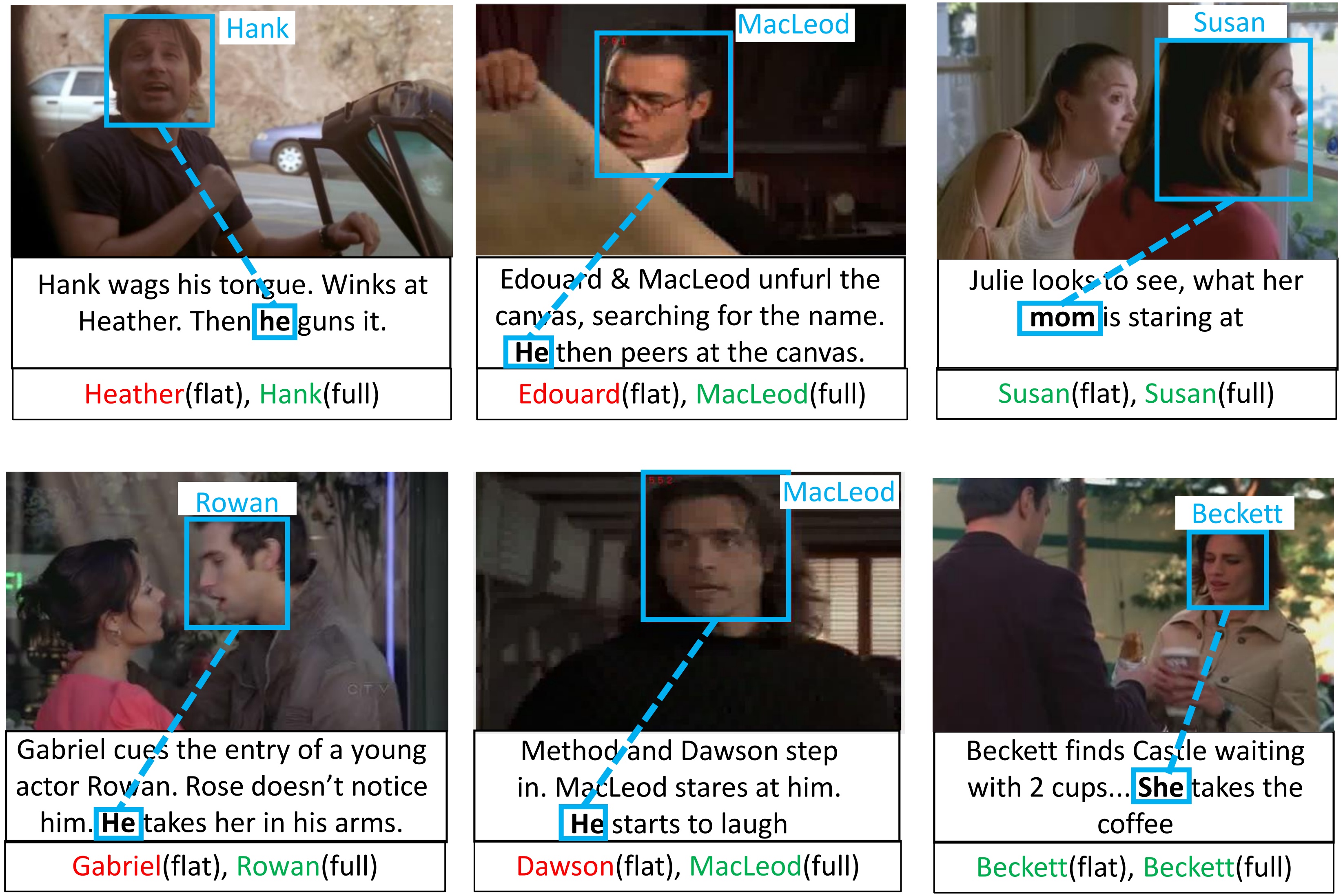
- Mean average-precision (mAP) scores for person name assignment

Results on coreference resolution

Set	Dev.	Test
CoreNLP [10]	52.01 %	40.59 %
Hagh. [2] modified	51.36 %	38.61 %
Our flat	54.85 %	44.22 %
Our+uni.	55.50 %	48.84 %
Our+rand.	55.11 %	45.54 %
Our (full)	**57.31 %**	**52.15 %**

- Accuracy of mention associated with the correct person name

Qualitative results



Conclusion

- We tackle jointly a vision and NLP problem and show improvement on both sides when combined
- Future work:
  - Simplified our model?
  - How to take into account actions? Or could this be used to learn more principled action “classifier”?

Efficient Image and Video Co-localization with Frank-Wolfe Algorithm

With Kevin Tang and Li Fei-Fei

ECCV 2014

Problem statement

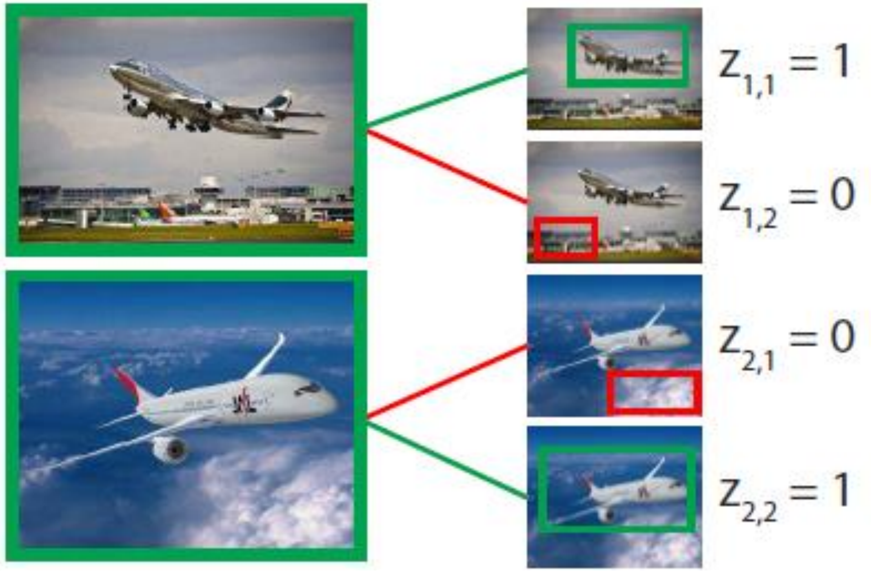
Image Co-localization | Video Co-localization

- A set of image/video containing the **same class** of object
- With no further supervision, **localize** all the instances

Our approach

- Select best bounding box per frame/image
- Our approach relies on a weakly supervised formulation introduced in Bach and Harchaoui (2008, NIPS)
- We show how to efficiently deal with lot of videos

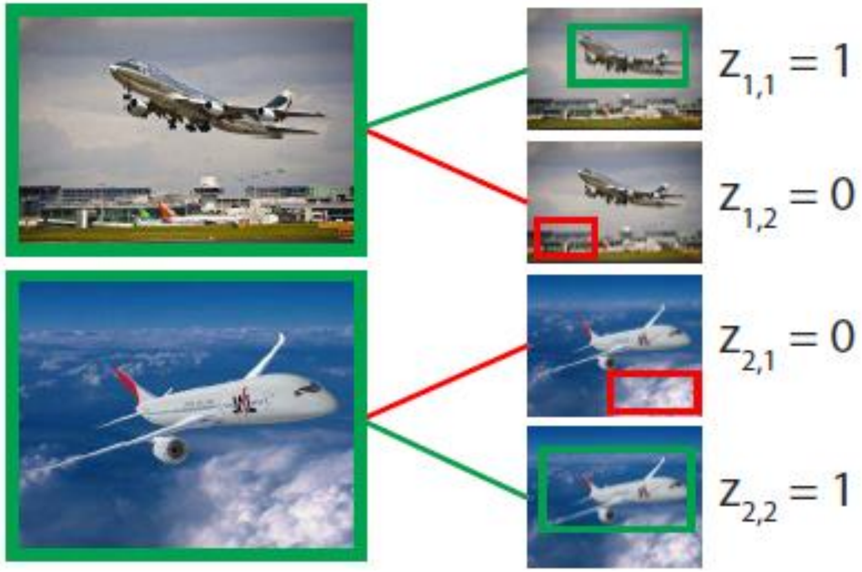
Discriminative model



- A box discriminability term:

$$\min_{\substack{w \in \mathbb{R}^d, \\ c \in \mathbb{R}}} \frac{1}{n_b} \sum_{j=1}^{n} \sum_{k=1}^{m} ||z_{j,k} - w x_{j,k}^{box} - c||_2^2 + \frac{\kappa}{d} ||w||_2^2,$$

Discriminative model

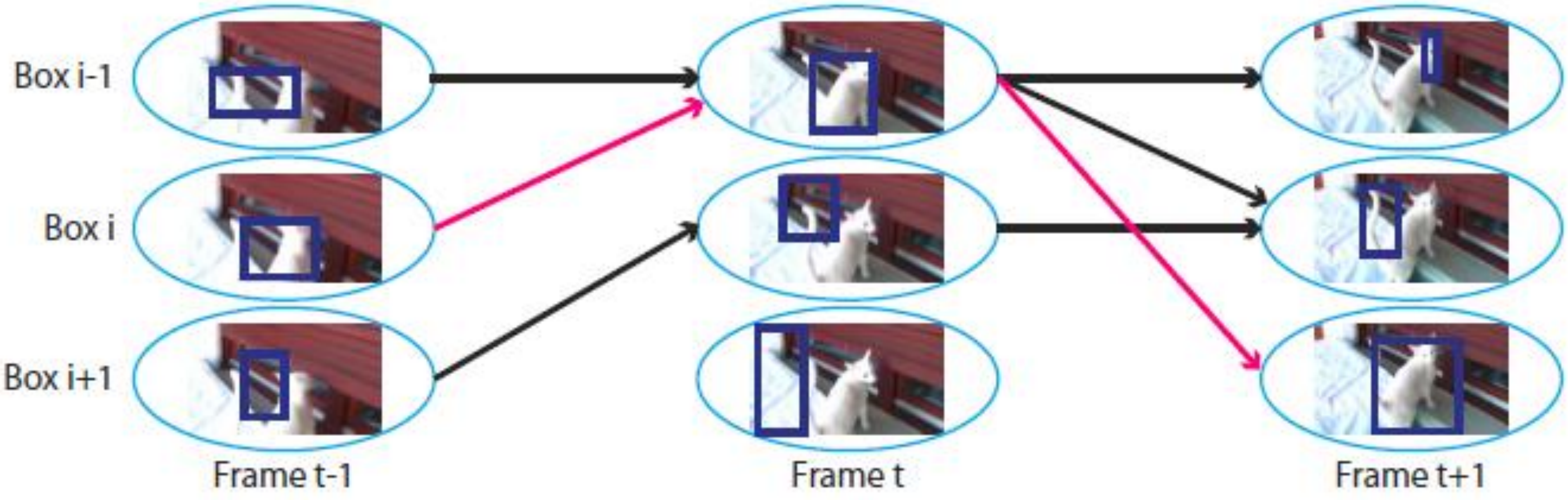


- Leading the quadratic convex function over z:

$$z^T A_{box} z,$$

Where A_{box} is a semi definite positive matrix (see Bach and Harchaoui, 2008)

Time consistency



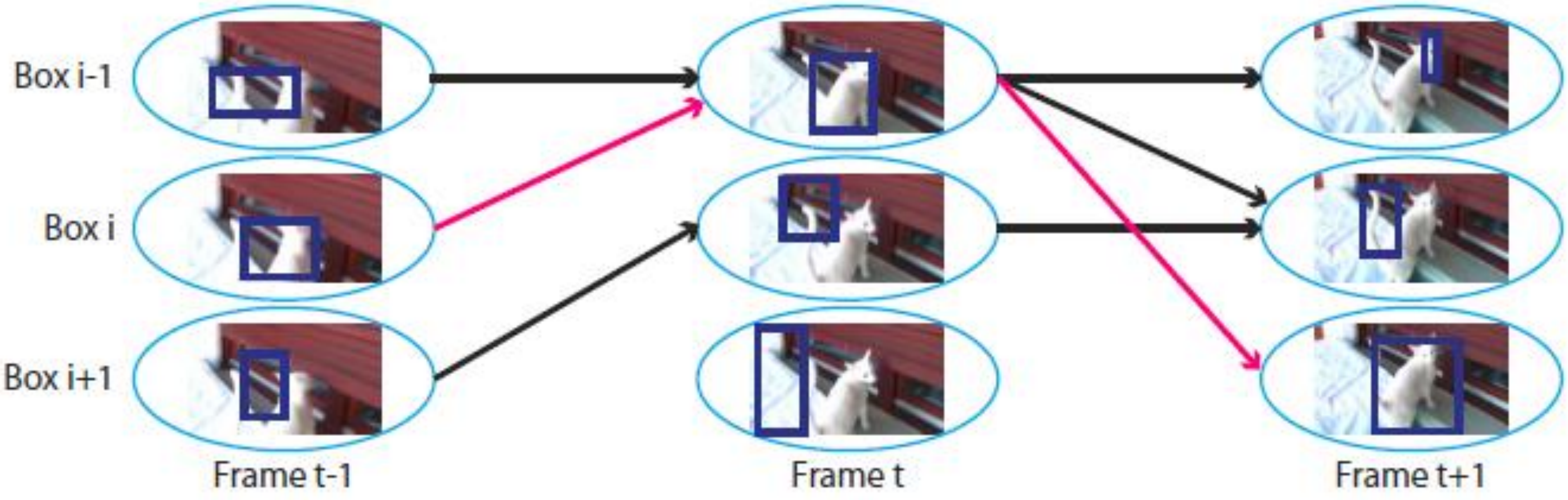
- A time consistency similary term:

$$s_{\text{temporal}}(b_i, b_j) = \exp\left(- \|b_i^{center} - b_j^{center}\|_2 - \left\| \frac{|b_i^{area} - b_j^{area}|}{\max(b_i^{area}, b_j^{area})} \right\|_2 \right)$$

On which we build a Laplacian matrix:

$$L_{box} = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$$

Time consistency

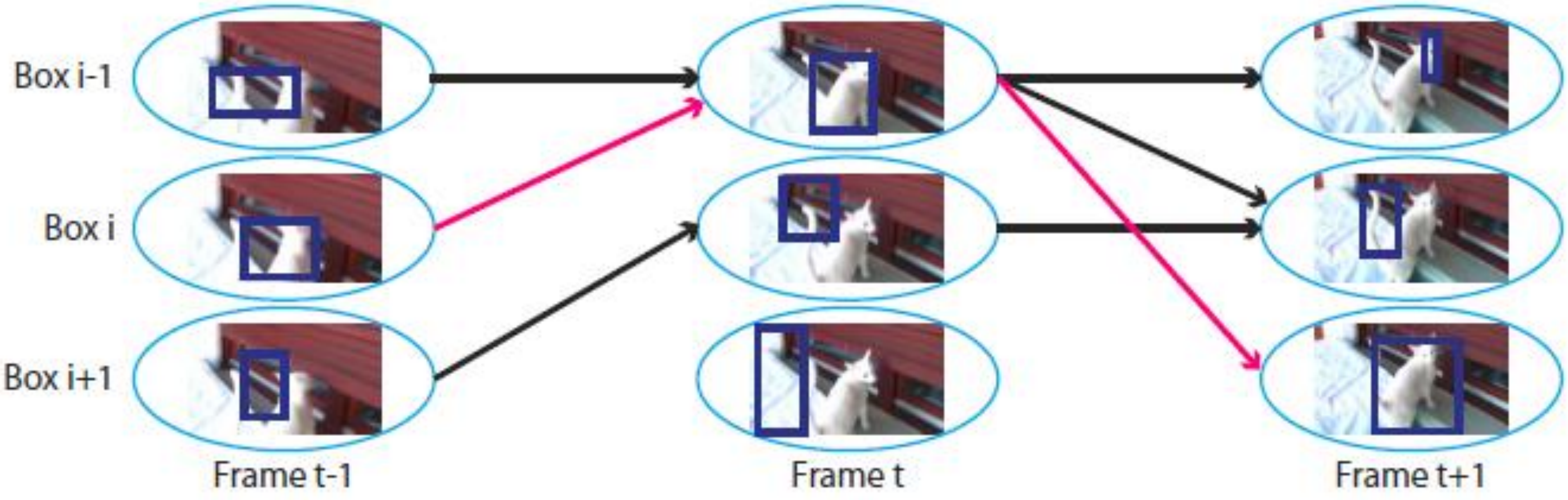


- Leading to another quadratic convex function:

$$z^T L_{box} z$$

Since a Laplacian matrix is sdp.

Time consistency



- We have additional flow constrains to encourage smooth solutions:

$$\forall V_i \in \mathcal{V},\ \forall k \in V_i,\ z_k = \sum_{l \in p(k)} y_{i,l,k} = \sum_{l \in c(k)} y_{i,k,l},$$

Overall problem

$$
\begin{aligned}
\underset{z,y}{\text{minimize}} \quad & z^T(L + \mu A + \mu_t U)z - z^T \lambda \log(m) \\
\text{subject to} \quad & z \in \{0,1\},\ y \in \{0,1\}, \\
& \forall I_j \in \mathcal{I} : \sum_{k=1}^{m} z_{j,k} = 1, \\
& \forall V_i \in \mathcal{V},\ \forall k \in V_i,\ z_k = \sum_{l \in p(k)} y_{i,l,k} = \sum_{l \in c(k)} y_{i,k,l},
\end{aligned}
$$

- Non-convex because of the discrete constraints
- Relax {0,1} to [0,1] => a convex problem
- **Problem**: Very large number of variables and constraints
- Standard solver are inefficient: O(N^3)
- **Solution:** Frank-Wolfe (FW) algorithm

Frank-Wolfe algorithm

- To minimize a function f over the convex set D, the FW algorithm solves at each iteration the following linear problem (LP):

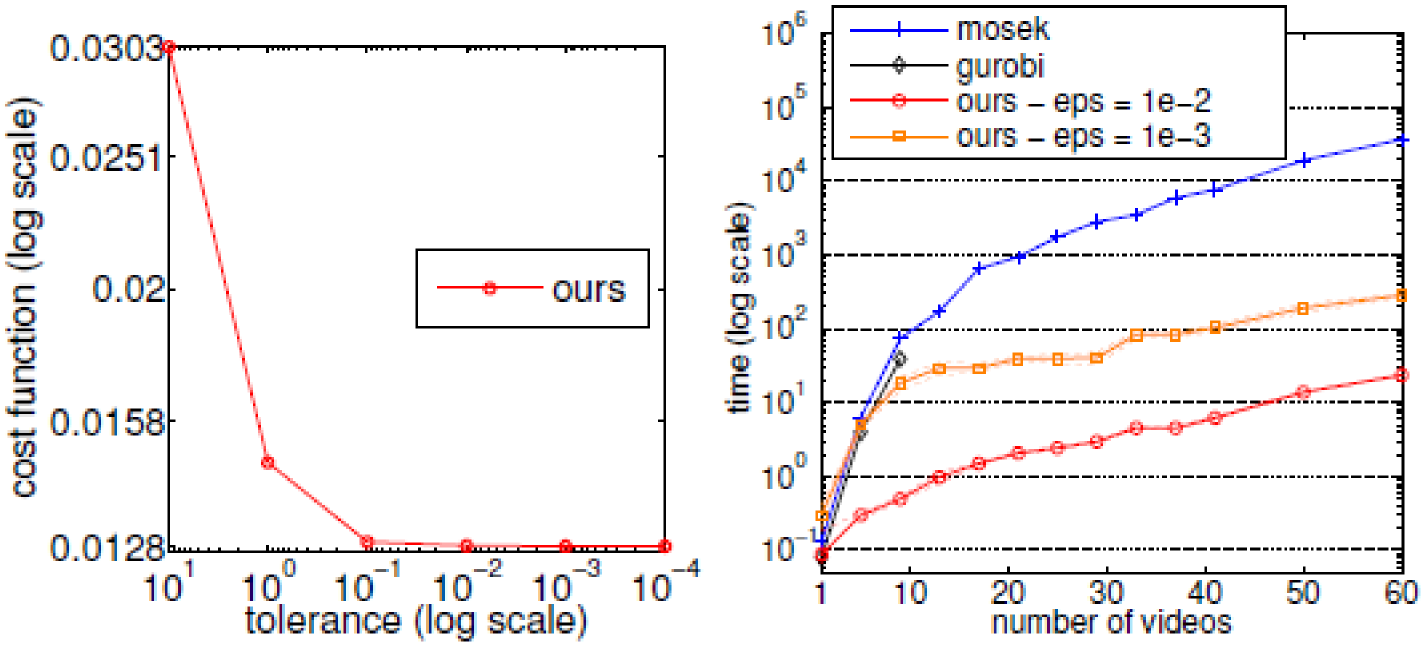
$$\begin{aligned} \underset{y}{\text{minimize}} \quad & y^T \nabla f(z_{k-1}) \\ \text{subject to} \quad & y \in \mathcal{D}. \end{aligned}$$

- In our case, this LP can be solved **efficiently** using a **shortest-path algorithm** for videos and a max function for the images

Related work

- This idea was used recently in other works:
  - Bojanowski *et al.* (ECCV, 2014) for action recognition in videos
  - Chari *et al.* (Arxiv, 2014) for multi-object tracking

Results: speed comparison



- For 80 videos, the FW algorithm takes 7 minutes
- We run >1000x faster than standard QP solvers

Results

Method	aeroplane	bird	boat	car	cat	cow	dog	horse	motorbike	train	Average
[37]	**51.7**	17.5	**34.4**	34.7	22.3	17.9	13.5	26.7	**41.2**	25.0	28.5
Our method (image)	18.36	19.35	28.57	32.97	32.77	25.68	38.26	30.14	15.38	21.43	26.29
Our method (image) w/ smoothing	21.26	21.51	30.95	36.26	35.29	25.68	38.26	35.62	15.38	23.21	28.34
Our method (video)	25.12	**31.18**	27.78	**38.46**	**41.18**	**28.38**	33.91	35.62	23.08	25.00	**30.97**

- Results on Youtube-Object dataset
- % of correct box following Pascal measure (inter/union > 50%)
- Small gain (<3%) over [37]
- Reason: Not enough videos (at most 80 per class)?

Results

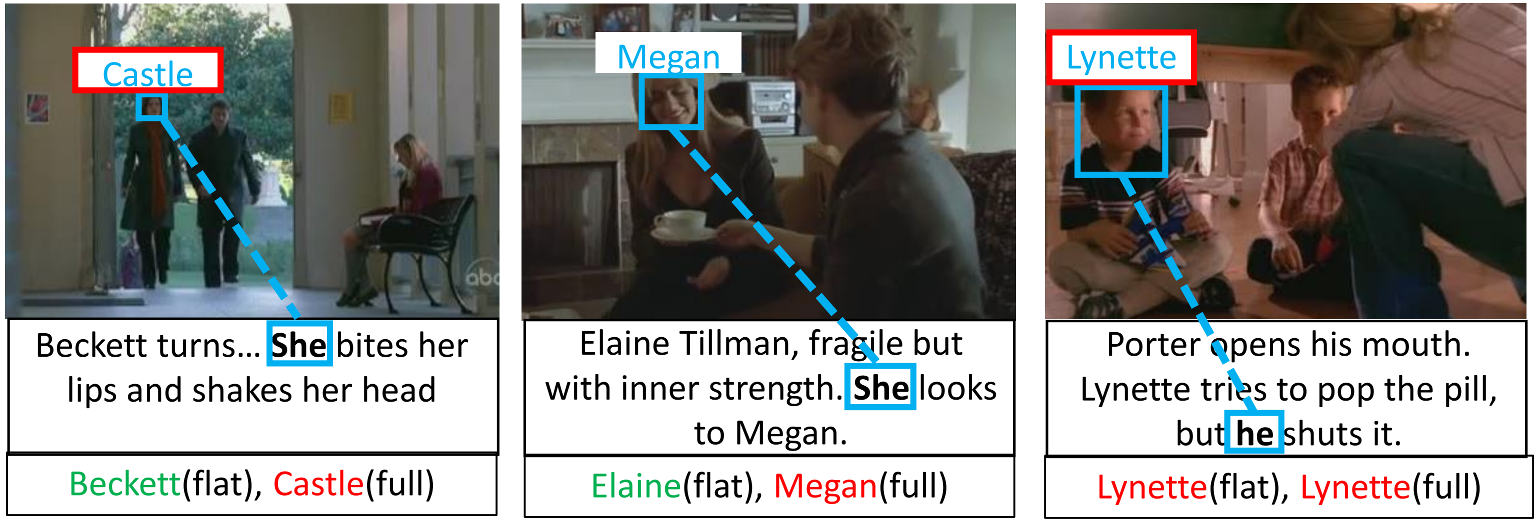
Qualitative comparison between our image model (red) and our video one (green)

Conclusion

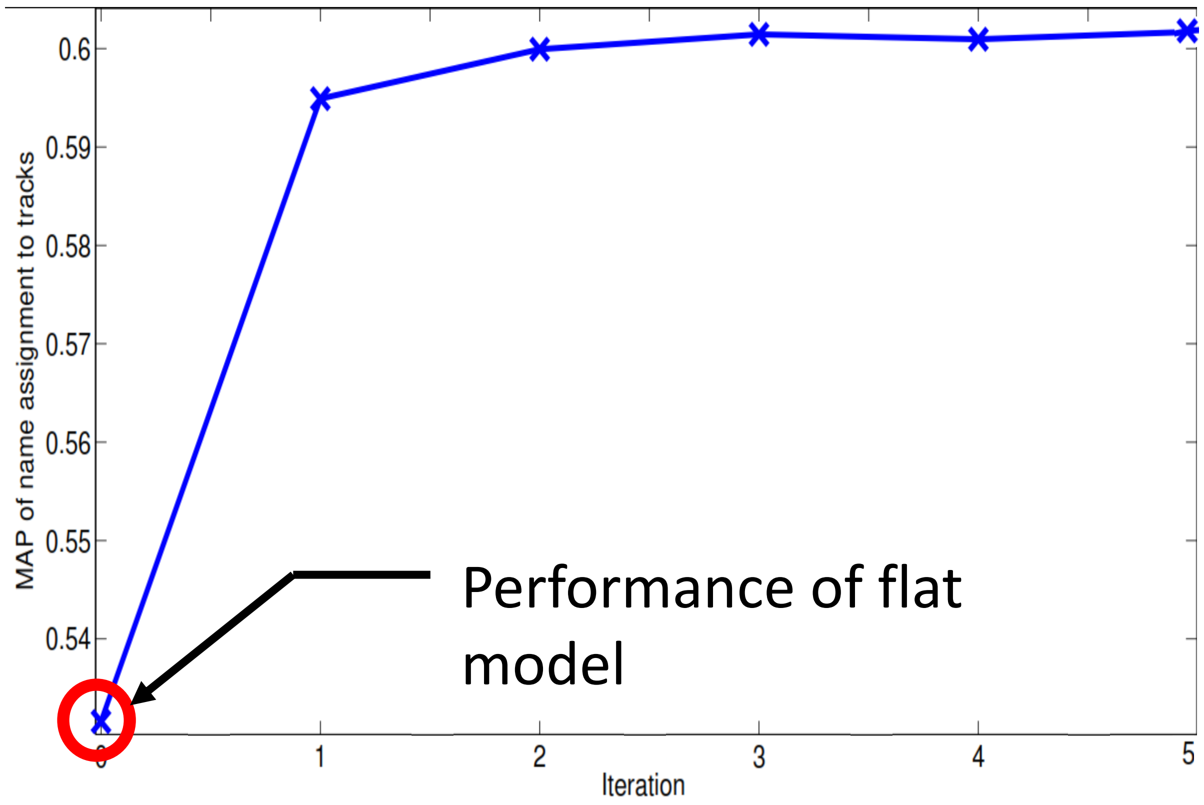
- We show an efficient algorithm for weakly supervised problem in videos
- Relatively small gain in localization performance

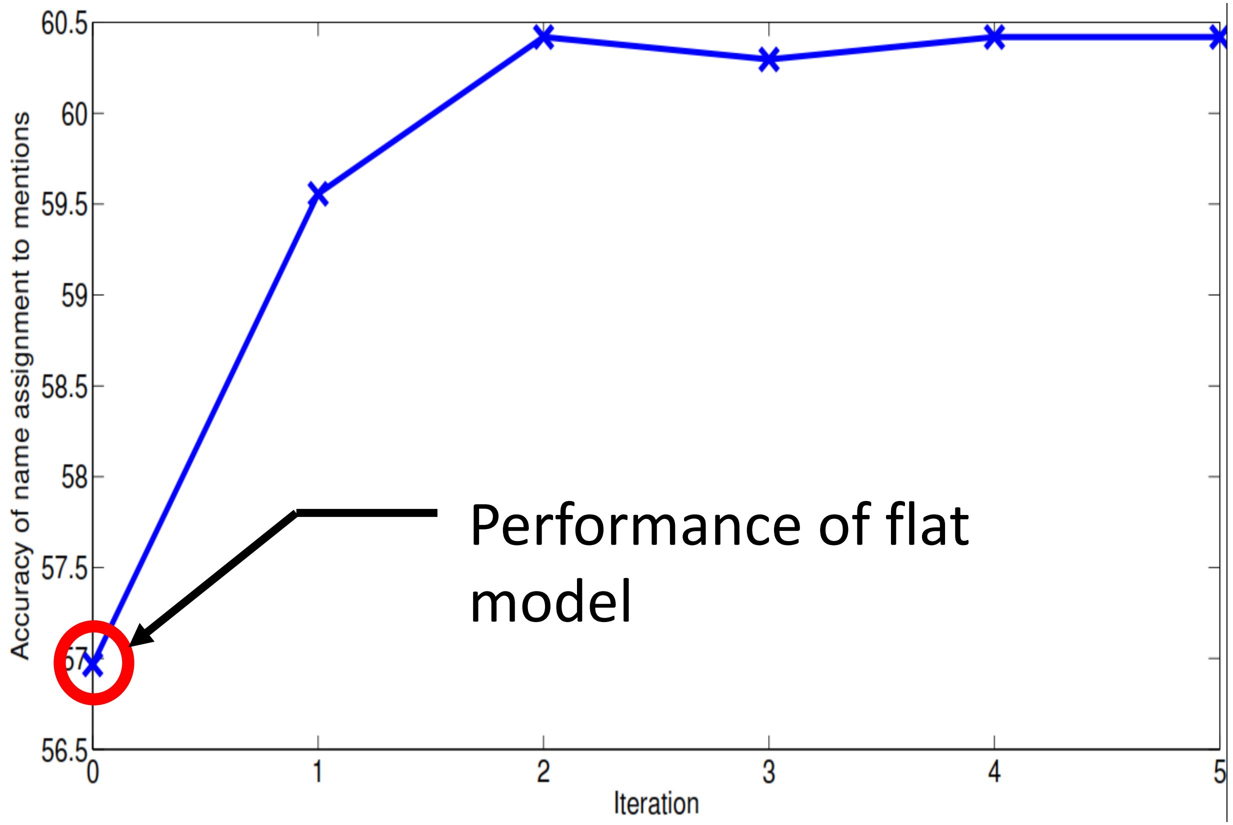
Thank you.

Failure cases



Performances with number of iterations





Results

Method	aeroplane	bird	boat	car	cat	cow	dog	horse	motorbike	train	Average
Video only	25.12	31.18	27.78	**38.46**	41.18	28.38	33.91	35.62	**23.08**	25.00	30.97
Joint Image+Video	**27.54**	**33.33**	27.78	34.07	**42.02**	28.38	**35.65**	35.62	21.98	25.00	**31.14**

- Surprisingly, adding images gives only a marginal boost…