Finding Actors and Actions in Movies

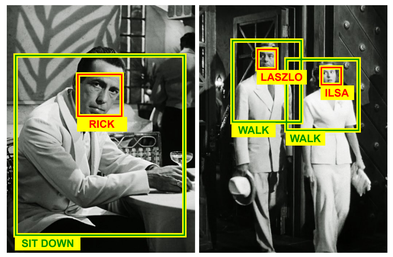
Piotr Bojanowski, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid and Josef Sivic.

INRIA / ENS / CNRS

Goal

- Recognize **people** and their **actions** in video.
- Use **weak supervision** derived from **video scripts**.
- Adress challenges of :
  - temporal localization,
  - spatial localizaion,
  - visual variability.

Rick sits down again and stares off in their direction. Ilsa and Laszlo leave the cafe.



Contributions

- Joint model for weakly-supervised learning of actions and actors.
- Solution in terms of quadratic problem with linear constraints.
- Improved results for action and face recognition.

Overview

Input : video sequence with associated script.

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...

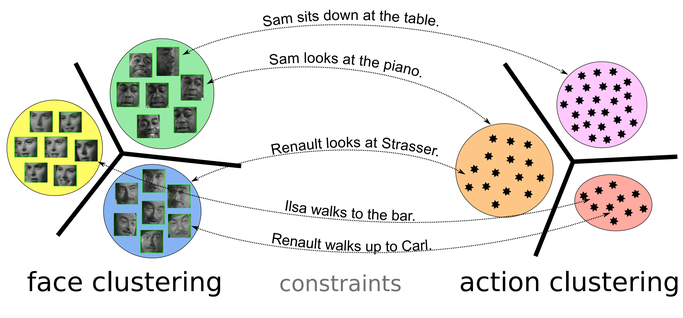
Output : tracks of people with names and action labels.

Overview of the proposed method :

1) Group faces and actions with similar visual appearence.

2) Assign names ("ILSA", etc...) and action labels ("WALK", etc...) to people.

3) Use joint action-name constraints derived from scripts.



Discriminative Clustering

optimizing over labels Z, classifiers f

$$\min_{Z,f} \frac{1}{N} \sum_{i \in I} \sum_{n \in \mathcal{N}_i} \ell\left(z_n, f(\phi(x_n))\right) + \Omega(f)$$

sum on bags; loss; classifier; regularizer; sum on samples; latent variable; features

Following [1], using linear classifiers and L2 regularization we get

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

Using the closed form solution of the Ridge Regression, solve for w and b

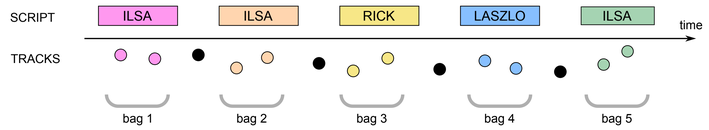
$$\min_Z \ Tr(ZZ^T A(X, \lambda_1))$$

Writing this problem for Persons and Actions jointly yields :

$$\min_{Z,T} Tr(ZZ^T A(X, \lambda_1)) + Tr(TT^T B(X, \lambda_2))$$

This recovers groups but the permutation of labels is not fixed!

Constraints



Z : indicator matrix such that $z_{np} = 1$ if sample n is of class p

At least one sample in bag i is of class p $\iff$ $\sum_{n \in \mathcal{N}_i} z_{np} \geq 1$

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam.

$\forall (p, a) \in \Lambda_i \ , \ \sum_{n \in \mathcal{N}_i} z_{np} \, t_{na} \geq 1$ ← QUADRATIC CONSTRAINTS

$\forall (p, \varnothing) \in \Lambda_i \ , \ \sum_{n \in \mathcal{N}_i} z_{np} \geq 1$

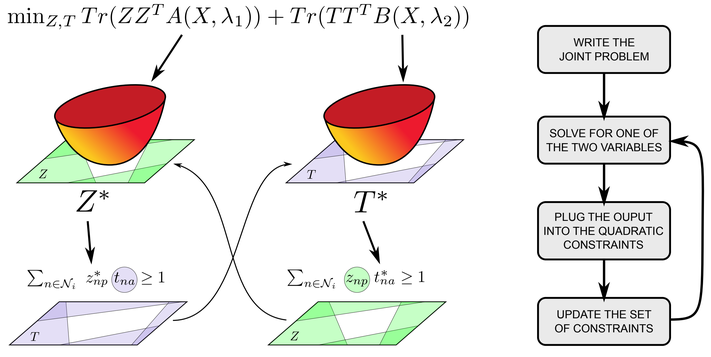
$\forall (\varnothing, a) \in \Lambda_i \ , \ \sum_{n \in \mathcal{N}_i} t_{na} \geq 1$ ← LINEAR CONSTRAINTS

Optimization

Quadratic cost under quadratic constraints $\quad \forall (p, a) \in \Lambda_i \ , \ \sum_{n \in \mathcal{N}_i} z_{np} \, t_{na} \geq 1$

Quadratic cost under linear constraints if we fix one of the variables

$$\min_{Z,T} Tr(ZZ^T A(X, \lambda_1)) + Tr(TT^T B(X, \lambda_2))$$



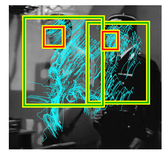
Results

Experiments on two movies : Casablanca, American Beauty

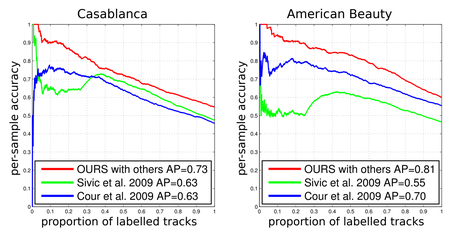
Features

- Detect Faces [4]
- Group Faces into tracks.
- Locate Facial Landmarks
- Warp to average location
- Re-locate landmarks
- Extract sifts
- Min-Min kernel [3]

- Given face tracks :
- Extrapolate face bounding box
- Compute dense trajectories [5]
- BOW inside bounding box



Results for faces



Evaluation of **character identification**.

Predict a character for every track.

Comparision with state of the art ([2], [3]) :

[2] : Defining a **loss on ambiguous labels**.

[3] : Relying on **speaker identification**.

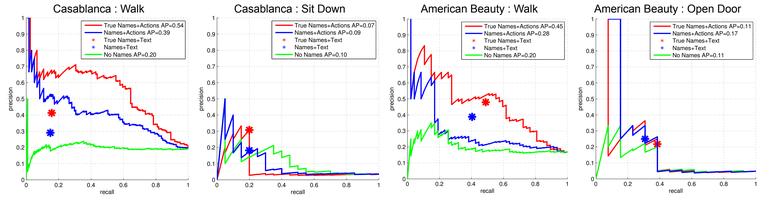
We outperform these methods.

Speaker identification performance is low.

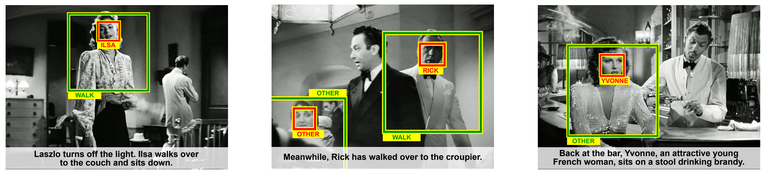
The hypothesis used in [2] isn't satisfied.

Top scored face tracks in Casablanca

Results on actions



Examples



Code and data available on-line : http://www.di.ens.fr/willow/research/actoraction/

References

[1] : F. Bach and Z. Harchaoui. Diffrac: a discriminative and flexible framework for clustering. In NIPS, 2007.

[2] : T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In CVPR, 2009.

[3] : J. Sivic, M. Everingham, and A. Zisserman. "who are you?" - learning person specific classifiers from video. In CVPR, 2009.

[4] : X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In CVPR, 2012.

[5] : H. Wang, A. Klaser, C. Schmid, and L. Cheng-Lin. Action Recognition by Dense Trajectories. In CVPR, 2011.