Emerging Topics in Human Activity Recognition

Michael Ryoo NASA Jet Propulsion Laboratory
Ivan Laptev INRIA
Greg Mori Simon Fraser University
Sangmin Oh Kitware

CVPR tutorial on 2014/06/23

Aaction Recognition with Bag-of-Features and Beyond

Ivan Laptev
INRIA Paris

CVPR tutorial on 2014/06/23

Computer vision grand challenge: Video understanding

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…

What are the challenges?

- **Large variations in appearance:** occlusions, non-rigid motion, view-point changes, clothing…

Action *Hugging*:



…

- **Manual collection of training samples is prohibitive:** many action classes, rare occurrence





…

- **Action vocabulary is not well-defined**

Action *Open*:




…

What are the challenges?

- **Large variations in appearance:** occlusions, non-rigid motion, view-point changes, clothing…

Action *Hugging*:

…

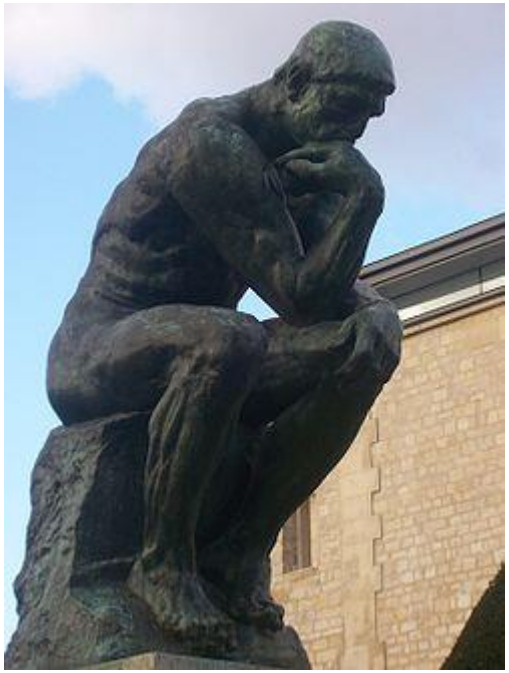
Part I

- Early methods
- Bag-of-features action classification

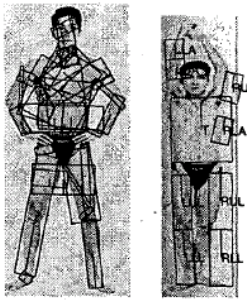
Part II

- Mid-level representations
- Temporal models of action recognition
- Action localization

Activities are characterized by pose

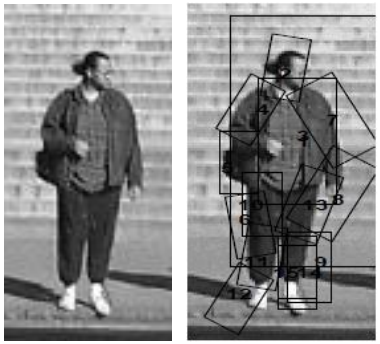


Pose estimation is difficult

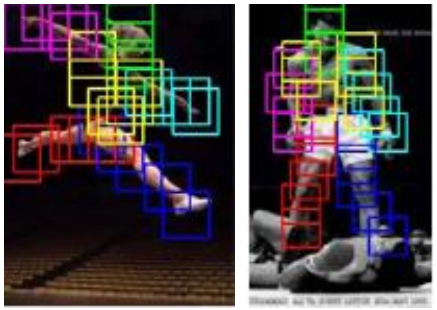


Finding People by Sampling Ioffe & Forsyth, ICCV 1999

Pictorial Structure Models for Object Recognition Felzenszwalb & Huttenlocher, 2000

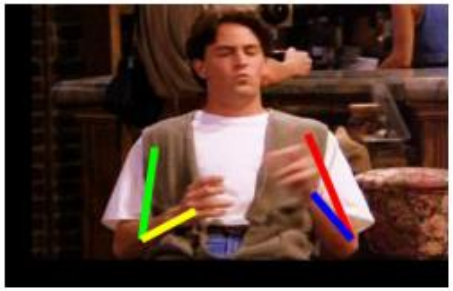


Learning to Parse Pictures of People Ronfard, Schmid & Triggs, ECCV 2002



Articulated pose estimation with flexible mixtures-of-parts. Y. Yang and D. Ramanan. CVPR 2011

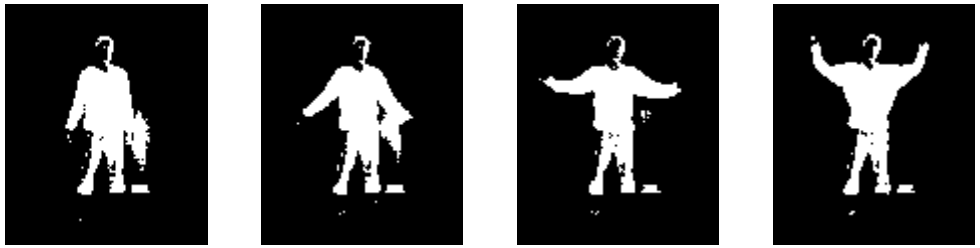
DeepPose: Human Pose Estimation via Deep Neural Networks, Toshev and Szegedy, CVPR 2014.



Mixing Body-Part Sequences for Human Pose Estimation, Cherian, Mairal, Alahari, Schmid, CVPR 2014.

Actions are more than poses

Appearance methods: Shape



[A.F. Bobick and J.W. Davis, PAMI 2001]
Idea: summarize motion in video in a
Motion History Image (MHI):

L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri.
Actions as spacetime shapes. 2007

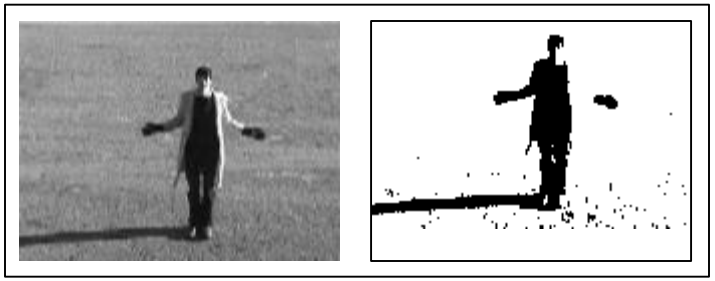
Appearance methods: Shape

Pros:

+ Simple and fast
+ Works in controlled settings

Cons:

- Prone to errors of background subtraction



Variations in light, shadows, clothing…

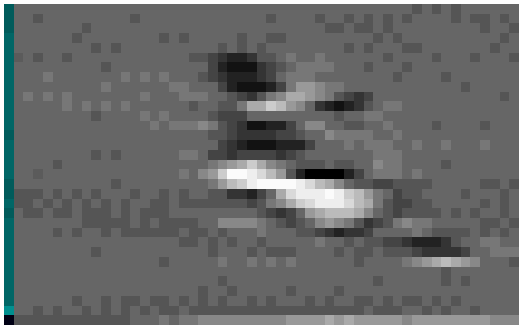
What is the background here?

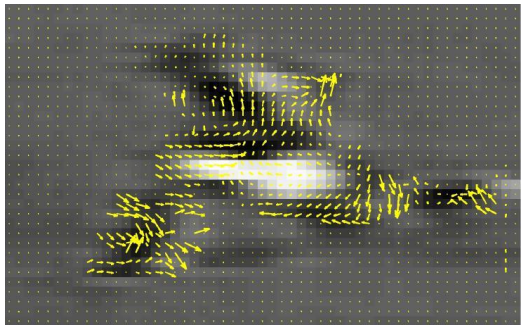
- Does not capture *interior* Structure and motion

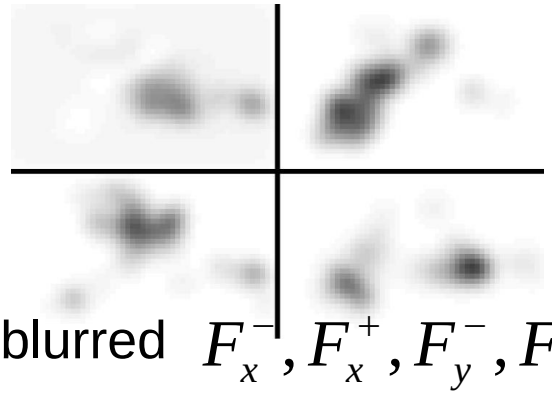
Silhouette tells little about actions

Appearance methods: Motion

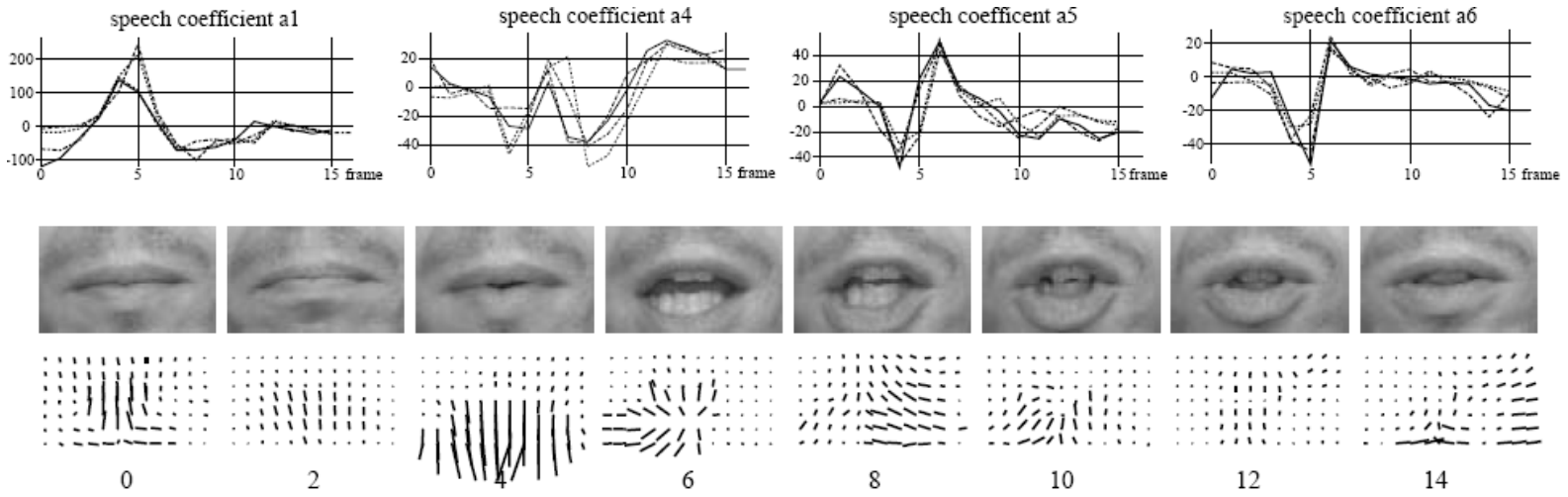
Recognizing action at a distance
A.A. Efros, A.C. Berg, G. Mori, and J. Malik., 2003.





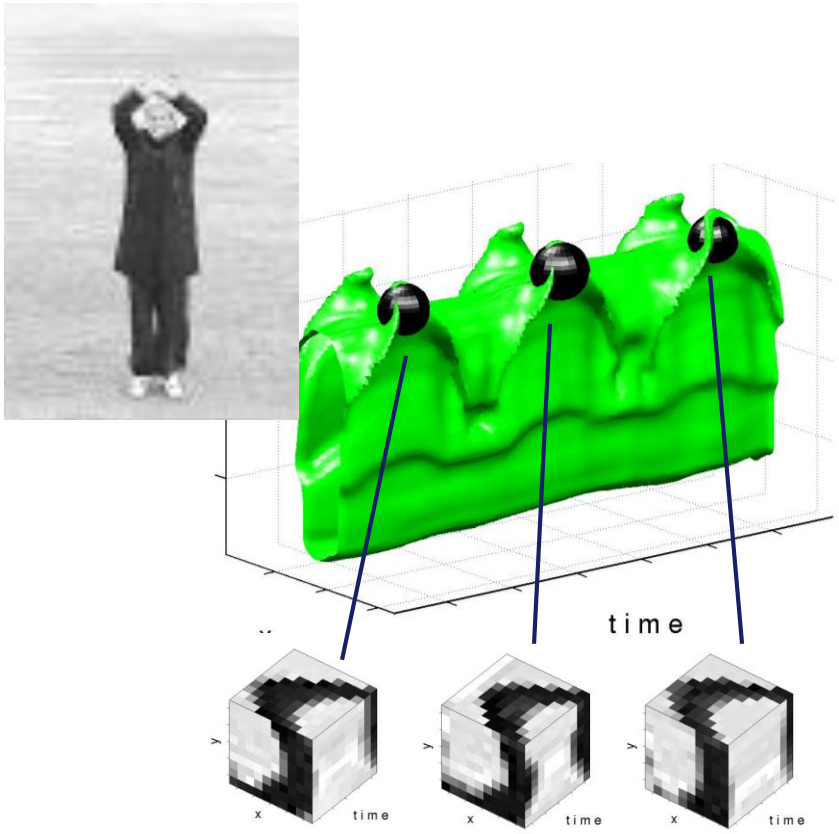


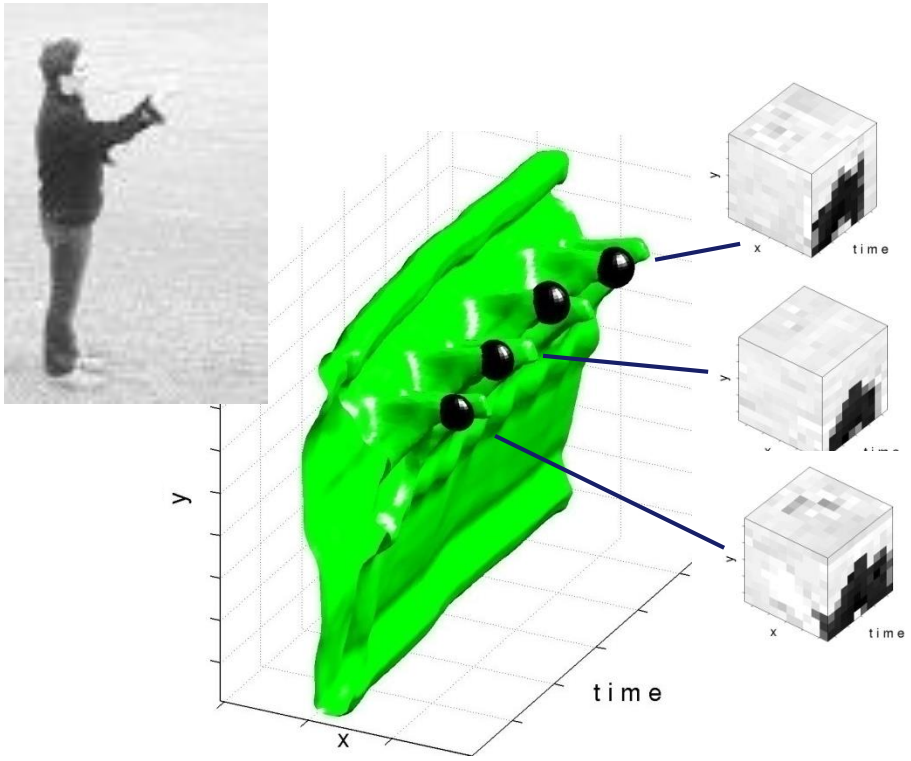
Learning Parameterized Models of Image Motion
M.J. Black, Y. Yacoob, A.D. Jepson and D.J. Fleet, 1997



Local feature methods

+ No segmentation needed
+ No object detection/tracking needed
- Loss of global structure





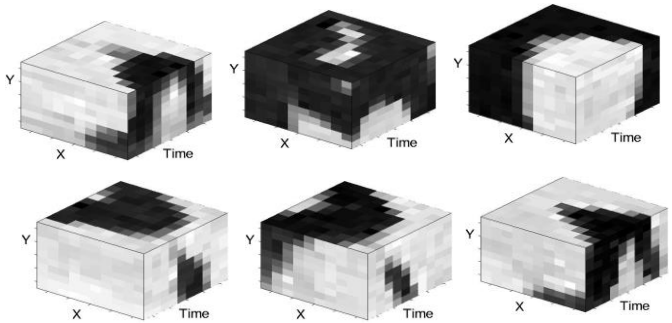
Local feature methods: Why working?

- Find similar events in pairs of video sequences

Bag-of-Features action recogntion

Extraction of Local features

space-time patches

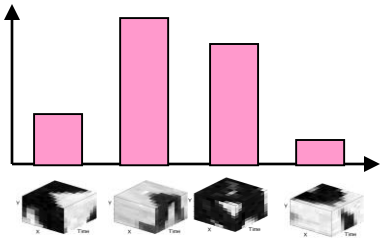


K-means clustering (k=4000)

Feature description

Feature quantization

Occurrence histogram of visual words



Non-linear SVM with χ^2 kernel

[Laptev, Marszałek, Schmid, Rozenfeld 2008]

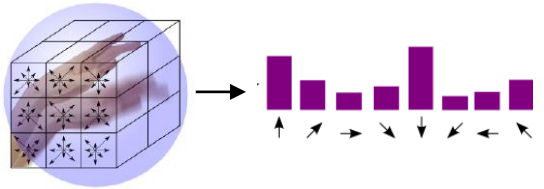
Action classification

Test episodes from movies “The Graduate”, “It’s a Wonderful Life”,
“Indiana Jones and the Last Crusade”

[Laptev et al. 2008]

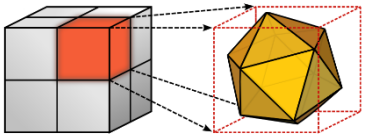
Alternative local video descriptors

- Laptev, Marszałek, Schmid and Rozenfeld, "Learning realistic human actions from movies", CVPR 2008.



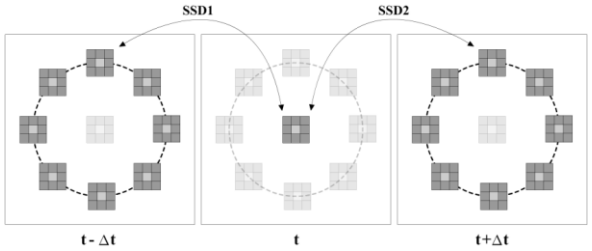
Histogram of optical flow (HOF)

- Klaser, Marszałek and Schmid, A Spatio-Temporal Descriptor Based on 3D-Gradients, BMVC 2008.

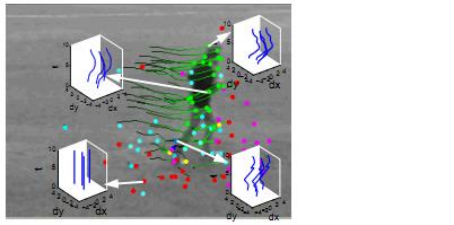


HOG3D

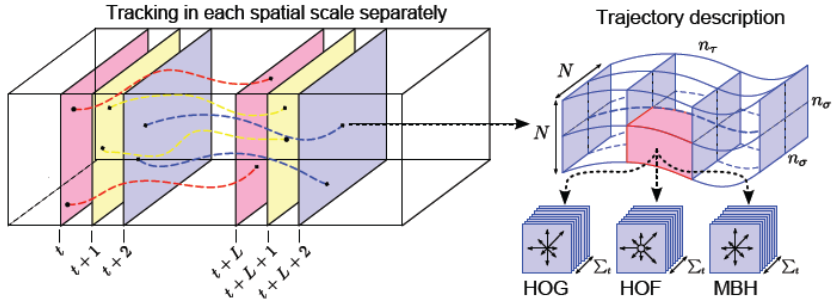
- Y. and L. Wolf, "Local Trinary Patterns for Human Action Recognition ", ICCV 2009, ECCV 2012



- P. Matikainen, R. Sukthankar and M. Hebert "Trajectons: Action Recognition Through the Motion Analysis of Tracked Features" ICCV VOEC Workshop 2009,

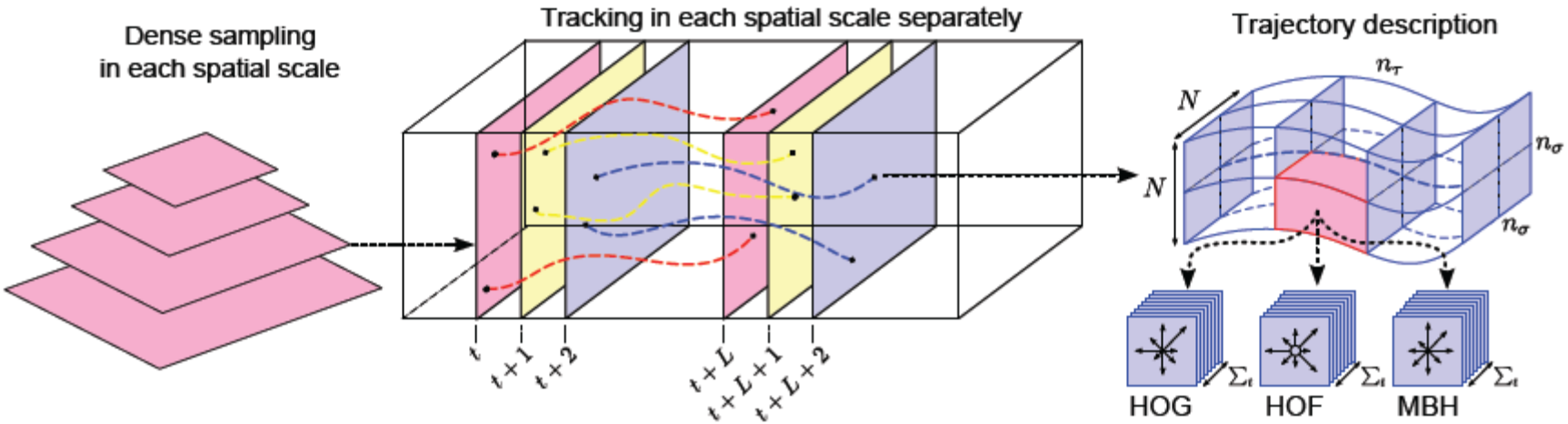


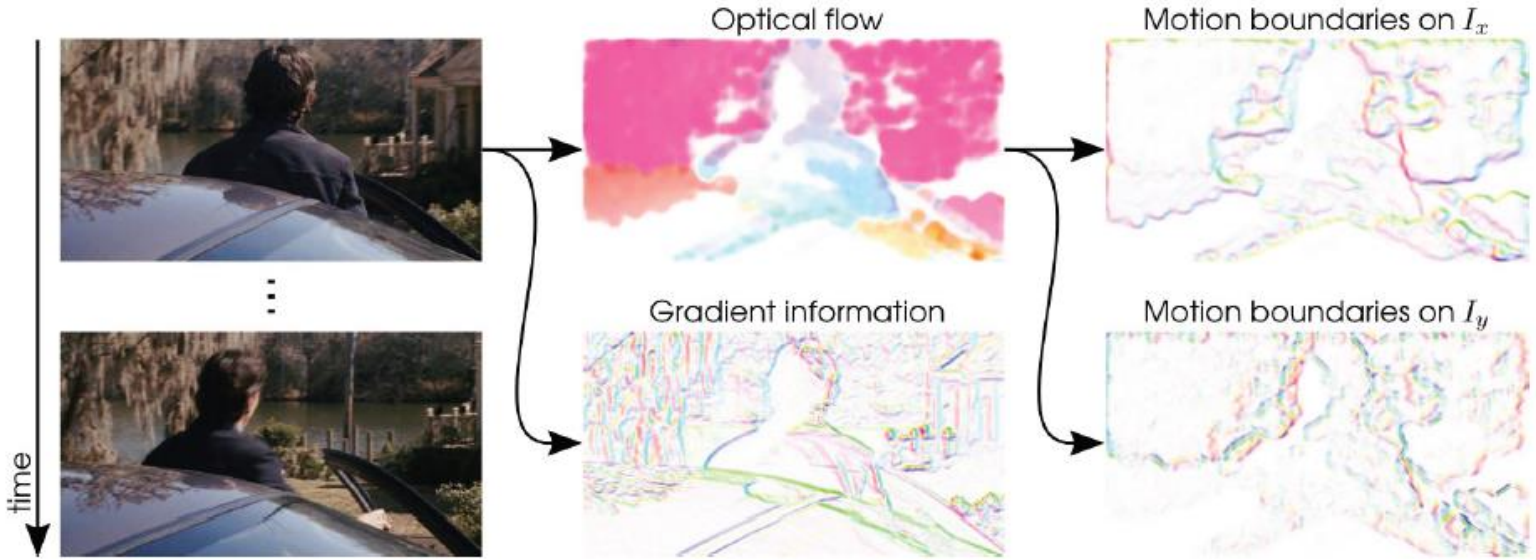
- H. Wang, A. Klaser, C. Schmid, C.-L. Liu, "Action Recognition by Dense Trajectories", CVPR 2011, IJCV 2013



Dense trajectory descriptors

[Wang et al. CVPR 2011]

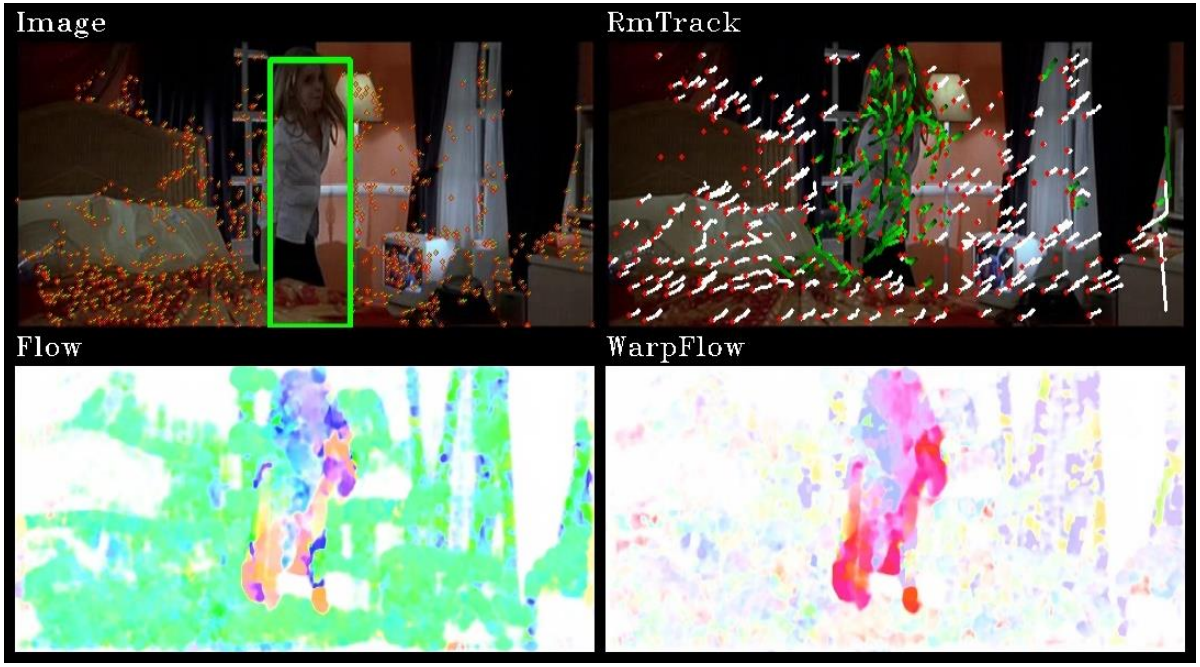




Improved trajectory descriptors

[Wang and Schmid ICCV 2013]

- Removes camera motion with Background motion estimation and person detection
- Uses Fisher Vector Encoding
- Similar ideas in Jain et al., "Better exploiting motion for better action recognition", In CVPR'13.



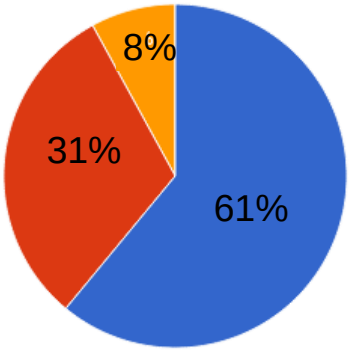
Hollywood2		HMDB51		Olympic Sports		UCF50	
Vig *et al.* [39]	59.4%	Sadanand *et al.* [32]	26.9%	Brendel *et al.* [4]	77.3%	Kliper-Gross *et al.* [17]	72.7%
Jiang *et al.* [15]	59.5%	Kliper-Gross *et al.* [17]	29.2%	Jiang *et al.* [15]	80.6%	Solmaz *et al.* [36]	73.7%
Mathe *et al.* [23]	61.0%	Jiang *et al.* [15]	40.7%	Gaidon *et al.* [12]	82.7%	Reddy *et al.* [31]	76.9%
Jain *et al.* [14]	62.5%	Jain *et al.* [14]	52.1%	Jain *et al.* [14]	83.2%	Shi *et al.* [34]	83.3%
Without HD	63.0%	Without HD	55.9%	Without HD	90.2%	Without HD	90.5%
With HD	**64.3%**	With HD	**57.2%**	With HD	**91.1%**	With HD	**91.2%**

Efficient features for action recognition

[Kantorov and Laptev CVPR 2014]

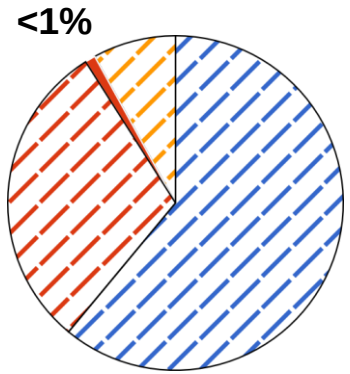
- Use sparse motion vectors from video compression.
- >100x speed-up of video feature extraction.

Time for video feature extracti on



Optical flow estimation
Descriptor aggregation
Tracking

Dense trajectories



This work

Hollywood2

	Acc.	Feat. (fps)	Quant. (fps)	Total (fps)
MF FLANN(4-32)	55.8%		52.4	40.0
MF VLAD(4)	56.7%	168.4	167.5	84.0
MF FV(32)	58.2%		40.9	32.9
DT	59.9%	1.2	5.1	1.0

UCF 50

	Acc.	Feat. (fps)	Quant. (fps)	Total (fps)
MF FLANN(4-32)	81.6%		52.4	48.1
MF VLAD(4)	80.6%	591.8	671.4	314.6
MF FV(32)	82.2%		171.3	132.8
DT	85.6%	2.8	5.1	1.8

What are the challenges?

- **Large variations in appearance:** occlusions, non-rigid motion, view-point changes, clothing…

Action *Hugging*:

…

Part I

- Early methods
- Bag-of-features action classification

Part II

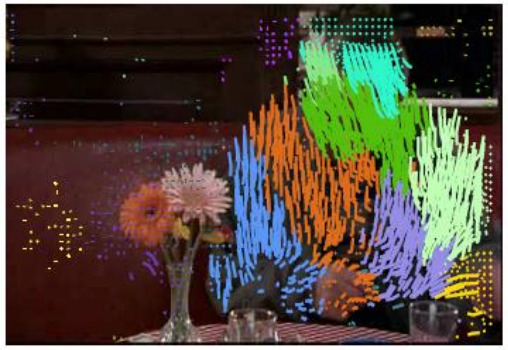
- Mid-level representations
- Temporal models of action recognition
- Action localization

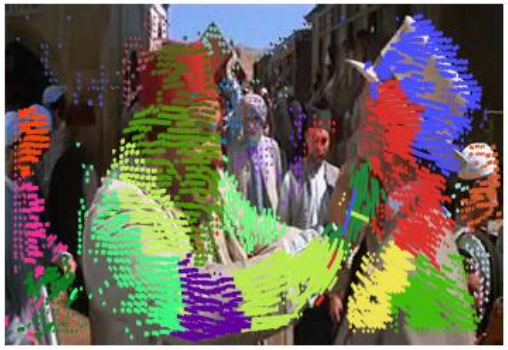
Mid-level action representations I

Discovering discriminative action parts from mid-level video representations
Raptis, Kokkinos, and Soatto, CVPR 2012.

- Groups point trajectories into tentative action parts by similarity in motion and appearance.

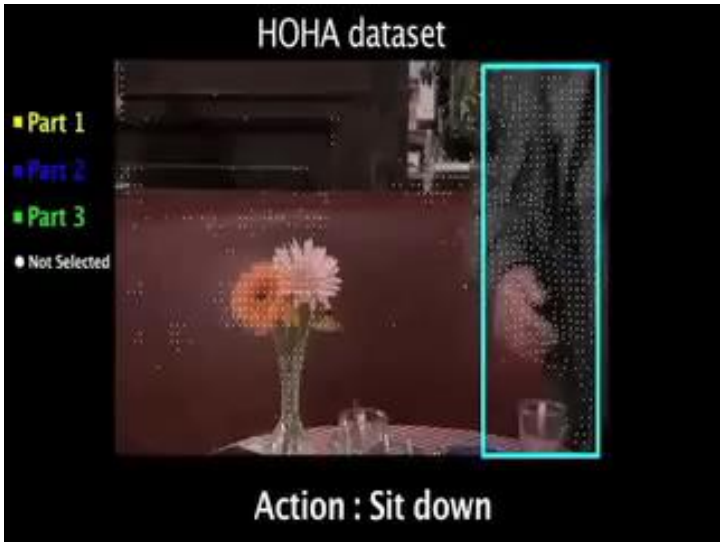
- Learns discriminative model with latent assignment of action parts.





⇨

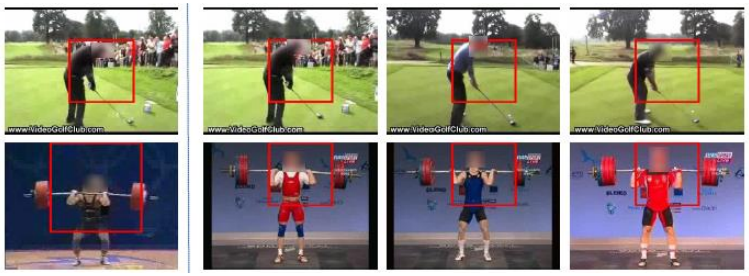
- Improved classification
- localization of discriminative action parts



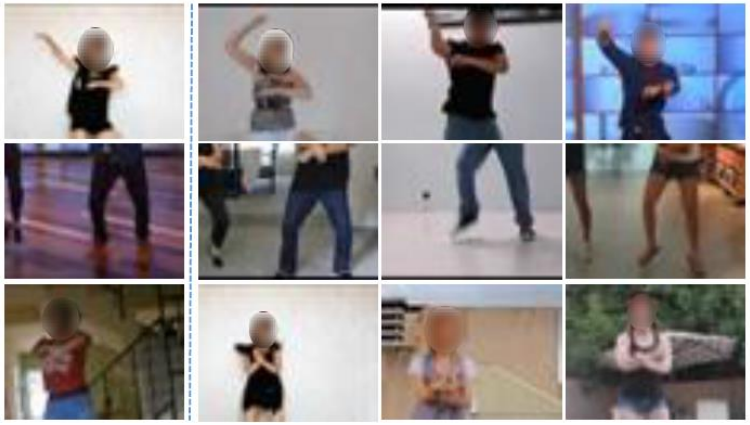
Mid-level action representations II

Representing Videos using Mid-level Discriminative Patches
Jain, Gupta, Rodriguez and Davis, CVPR 2013

- Mines discriminative and representative patches from action videos

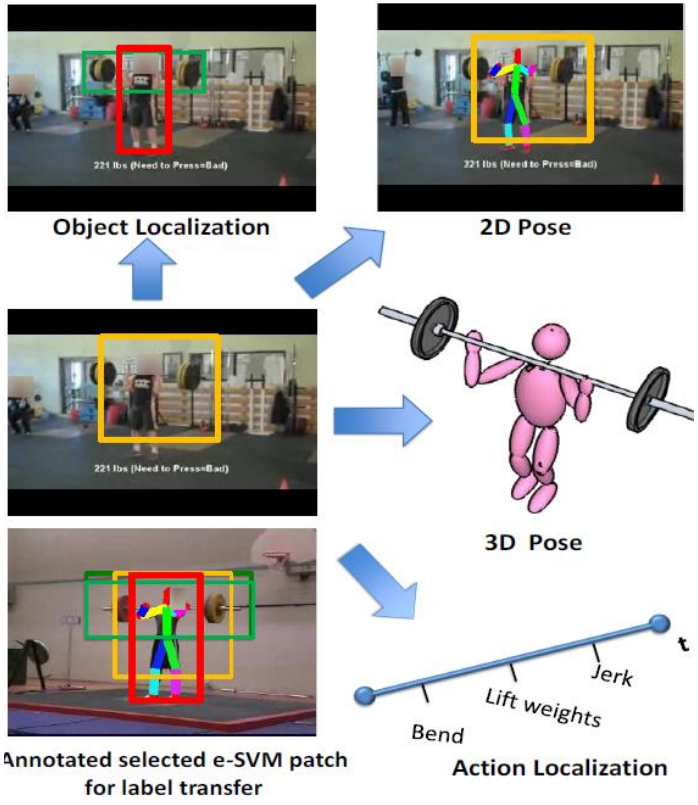


High-ranked discriminative spatio-temporal patches



Discriminative patches for “Gangnam” style

- Aligns patches and transfers additional annotation to test samples



Mid-level action representations

Conclusions:

- Enable part-based action analysis
  - localization of action parts
  - annotation transfer

- Current classification performance is inferior to the state of the art obtained with Fisher Vector + Improved dense trajectories.

- CNNs will take over soon?

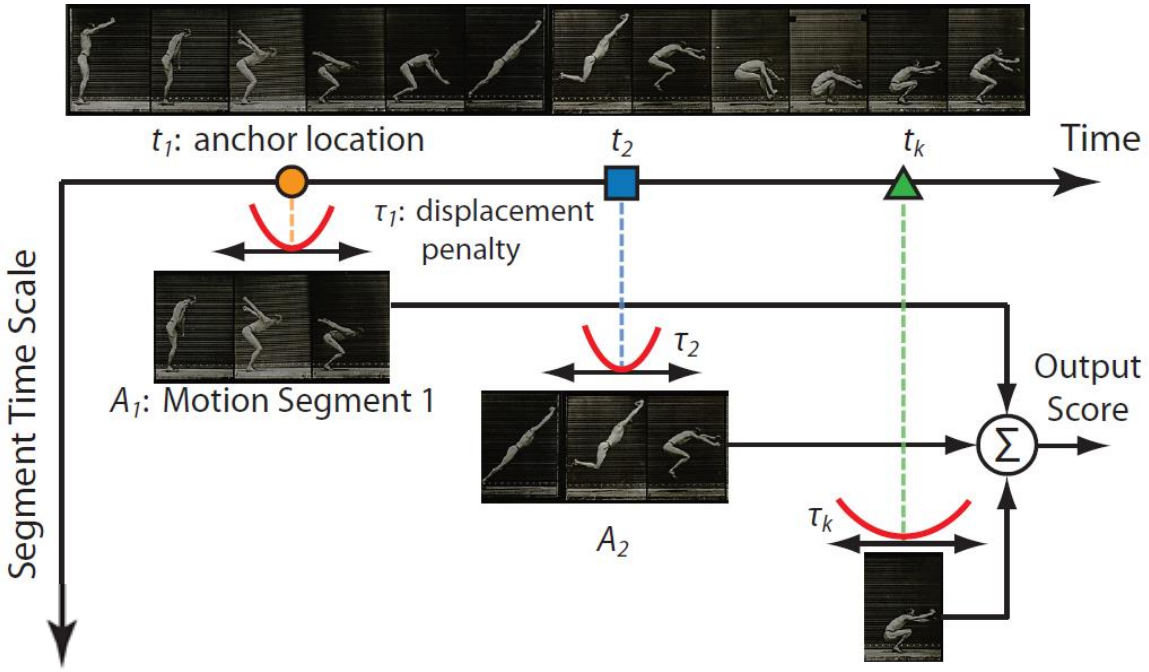
  Probably Yes, but:
  - What's a good network architecture?
  - Where to get sufficient training data?

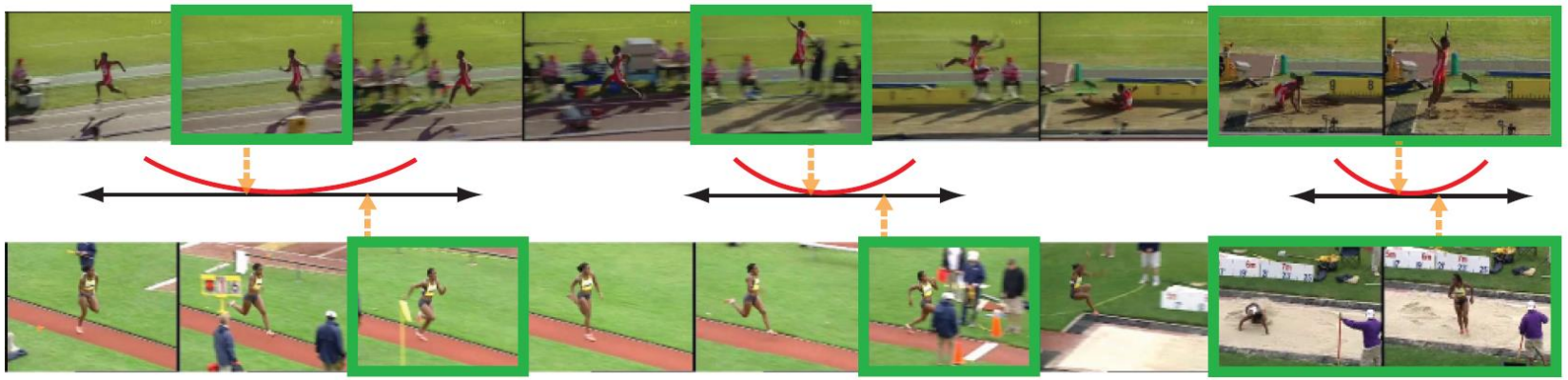
  Some work is on the way:
  Large-scale Video Classification using Convolutional Neural Networks Karpathy, Shetty, Toderici, Sukthankar, Leung and Fei-Fei, CVPR 2014 (Wednesday)

Temporal structure of actions I

Modeling Temporal Structure of Decomposable Motion Segments for Activity Classication, J.C. Niebles, C.-W. Chen and L. Fei-Fei, ECCV 2010

- Latent SVM model with temporal action parts.
- Enables temporal localization of action parts

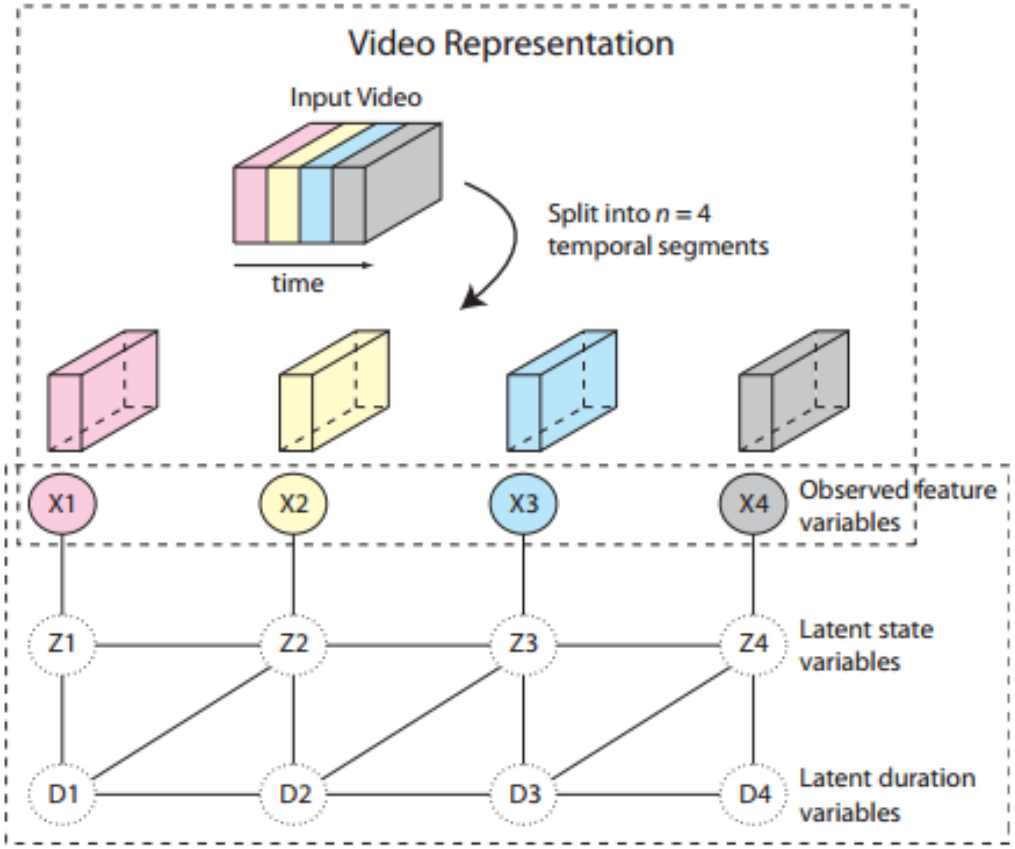




Temporal structure of actions II

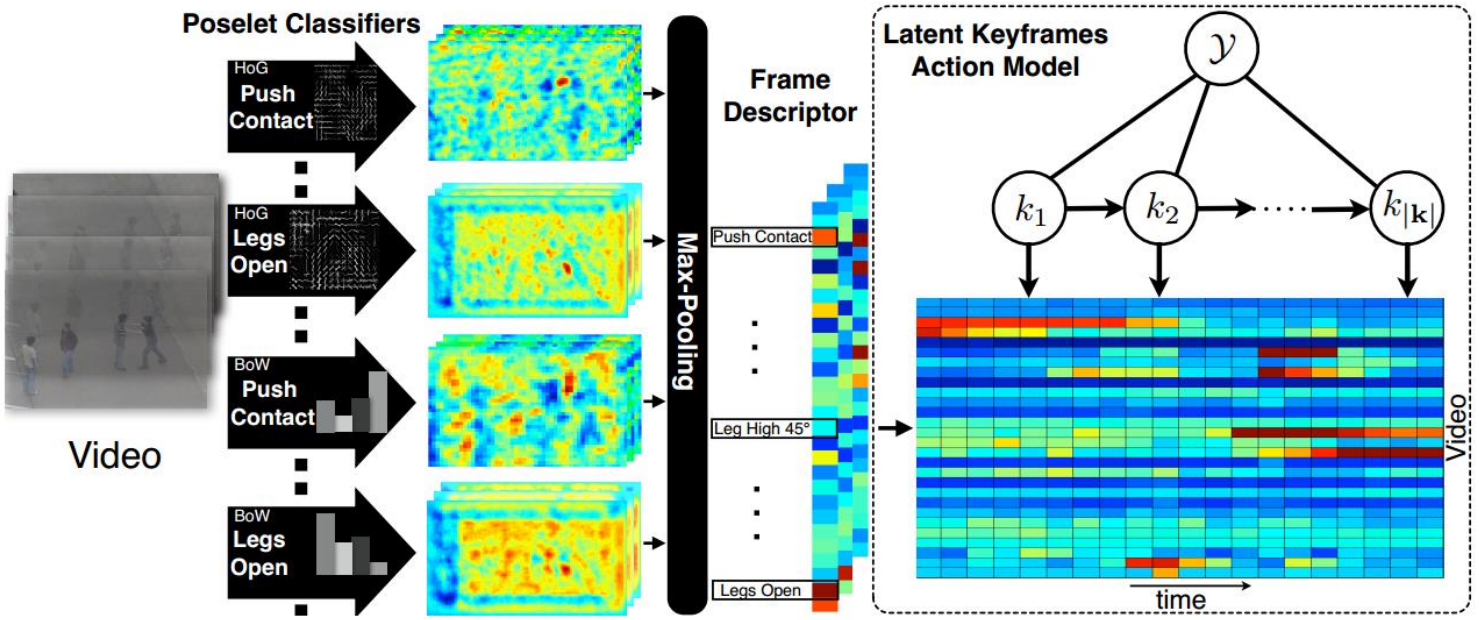
Learning Latent Temporal Structure for Complex Event Detection.
Kevin Tang, Li Fei-Fei and Daphne Koller, CVPR 2012

- Modeling of longer events such as Grooming an animal
- Discriminatively-trained Markov model
- Aims to infer and learn latent temporal structure of actions



Temporal structure of actions III

Poselet Key-framing: A Model for Human Activity Recognition.
Raptis and Sigal, In CVPR 20132012



Poselet activations

- Models actions as a sparse sequence of spatio-temporally localized key-frames

Temporal structure of actions IV

Scenario-based video event recognition by constraint flow.
Kwak, Han and Han, In CVPR 2011

- Models event by temporal logical composition of simple actions

Composite event	Scenario
Service[A]	Serve[A] $\sim$ Part[A, ball]
Stroke[A]	(Meet[A, ball] $<$ Part[A, ball]) $\wedge$ Swing[A]
TennisPlay†	Service[ser] $<$ (# | Service[ser]) $<$ (# | (Stroke[rec] $<$ Stroke[ser])$^+$) $<$ (# | Stroke[rec])

First service

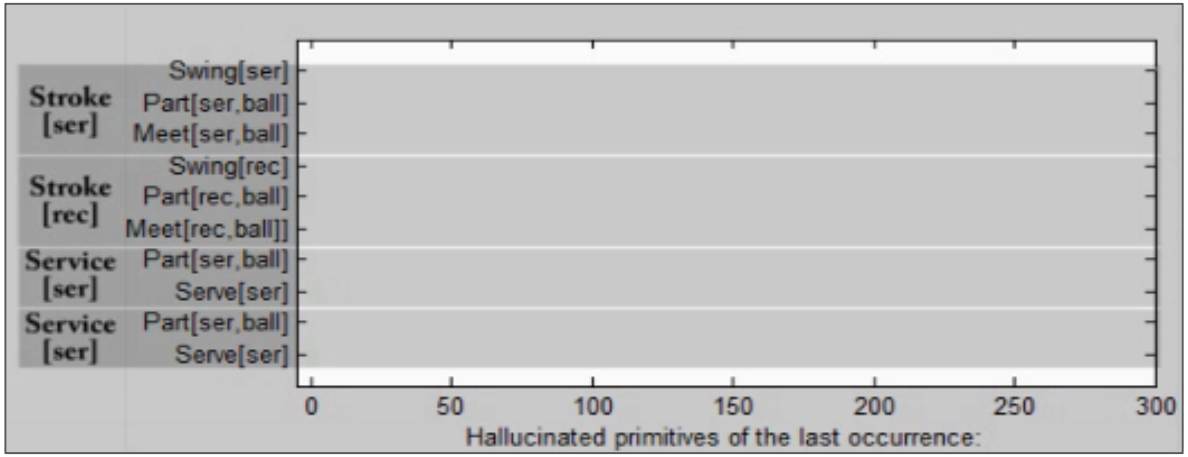
Second service
(may not happen)

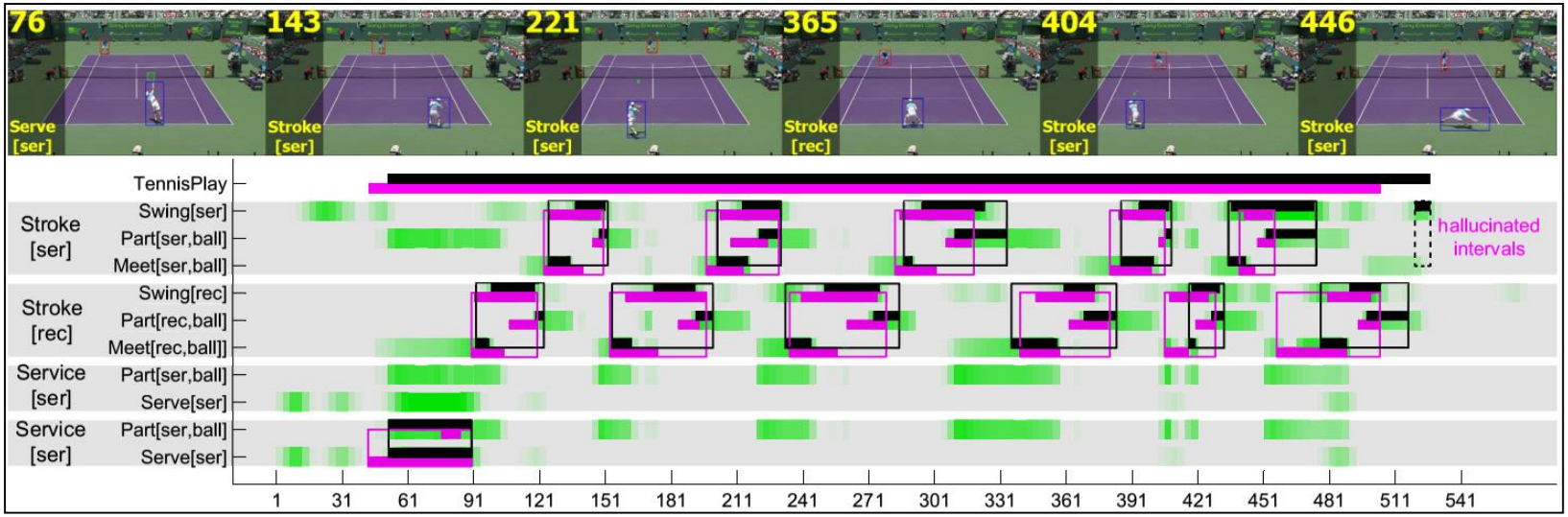
Unknown number of strokes

Temporal structure of actions IV

Scenario-based video event recognition by constraint flow.
Kwak, Han and Han, In CVPR 2011

Ours Groundtruth

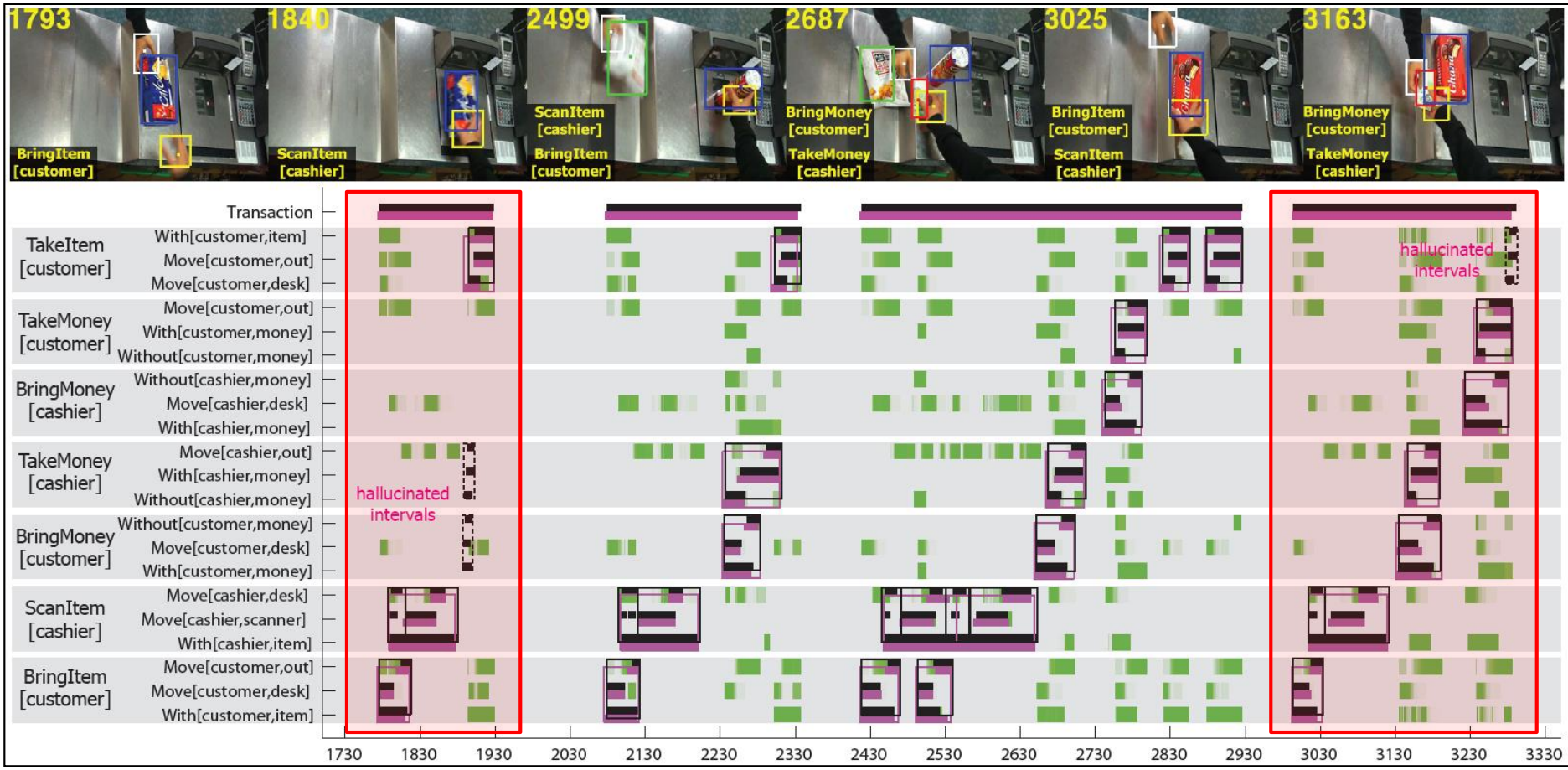




Temporal structure of actions IV

Scenario-based video event recognition by constraint flow.
Kwak, Han and Han, In CVPR 2011

Ours Groundtruth



"No pay"

"Missing"

Temporal activity models

Conclusions:

- Enable temporal analysis of sub-actions
- Enable modeling of long-term activities with variable structure of actions
- Current classification performance is inferior to the state of the art
- Open problems:

  How to automatically discover and learn scenarios?

  What are the meaningful action units for a target activity?

- Literature on action localization is relatively sparse
  => good area to make impact!
  See upcoming action recognition and localization challenge:
  THUMOS 2014 http://crcv.ucf.edu/THUMOS14/

References to *some* related work

Local features and bag-of-features representations

- Laptev and Lindeberg “Space-time interest points”, *In ICCV’03.*
- Schüldt et al., "Recognizing human actions: A local SVM approach", *In ICPR'04.*
- Dollar et al., “Behavior recognition via sparse spatio-temporal features”, *In VS-PETS’05*.
- Niebles et al., "Unsupervised learning of human action categories using spatial-temporal words", *In BMVC'06*.
- Laptev et al., “Learning realistic human actions from movies”, *In CVPR’08.*
- Wang et al., "Evaluation of local spatio-temporal features for action recognition", *In BMVC’09*.
- Wang, et al., " Action Recognition by Dense Trajectories", *In CVPR’11*.
- Jain et al., "Better exploiting motion for better action recognition", *In CVPR'13.*
- Wang and Schmid, "Action Recognition with Improved Trajectories“, *In ICCV'13.*
- Kantorov and Laptev, "Efficient feature extraction, encoding and classification for action recognition “, *In CVPR'14.*

Mid-level representations

- Liu et al., "Recognizing Human Actions by Attributes", *In CVPR'11*.
- Raptis et al., Discovering discriminative action parts from mid-level video representations." *In CVPR’12*.
- Sadanand and Corso. "Action bank: A high-level representation of activity in video." *In CVPR’12*.
- Jain et al., Representing Videos using Mid-level Discriminative Patches. *In CVPR’13*.

Temporal action models

- Ryoo and Agarwal, Recognition of composite human activities through context-free grammar based representation, *CVPR’06*
- Laxton et al., Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video*, CVPR’07.*
- Niebles et al., Modeling Temporal Structure of Decomposable Motion Segments for Activity Classication. *In ECCV 2010.*
- Kwak et al., Scenario-based video event recognition by constraint flow*, In CVPR 2011.*
- Khamis et al., Combining Per-Frame and Per-Track Cues for Multi-Person Action Recognition, *ECCV 2012.*
- Tang et al., Learning Latent Temporal Structure for Complex Event Detection. *In CVPR 2012.*
- Raptis and Sigal. Poselet Key-framing: A Model for Human Activity Recognition. *In CVPR 2013*
- Amer et al., Monte Carlo Tree Search for Scheduling Activity Recognition, *In ICCV 2013.*

What are the challenges?

- **Large variations in appearance:** occlusions, non-rigid motion, view-point changes, clothing…

Action *Hugging*:



…

- **Manual collection of training samples is prohibitive:** many action classes, rare occurrence





…

- **Action vocabulary is not well-defined**

Action *Open*:




…

What are the challenges?

- **Large variations in appearance:** occlusions, non-rigid motion, view-point changes, clothing…

Action *Hugging*:


…

- **Manual collection of training samples is prohibitive:** many action classes, rare occurrence




…

- **Action vocabulary is not well-defined**

Action *Open*:



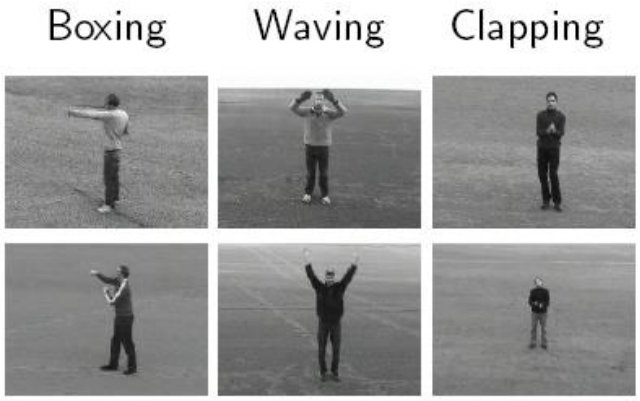
…

Where to get training data?

- Shoot actions in the lab

  KTH dataset
  Weizman dataset,…

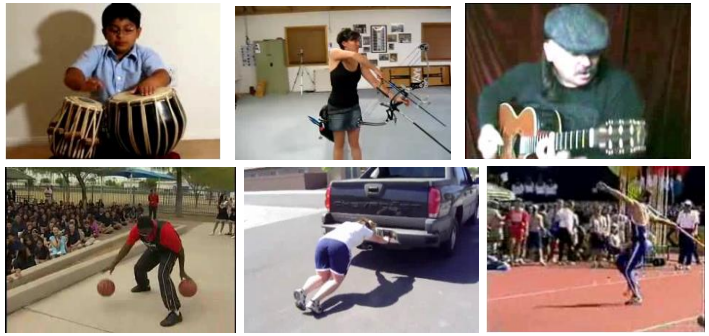
  ⇨ - Limited variability
  - Unrealistic



- Manually annotate existing content

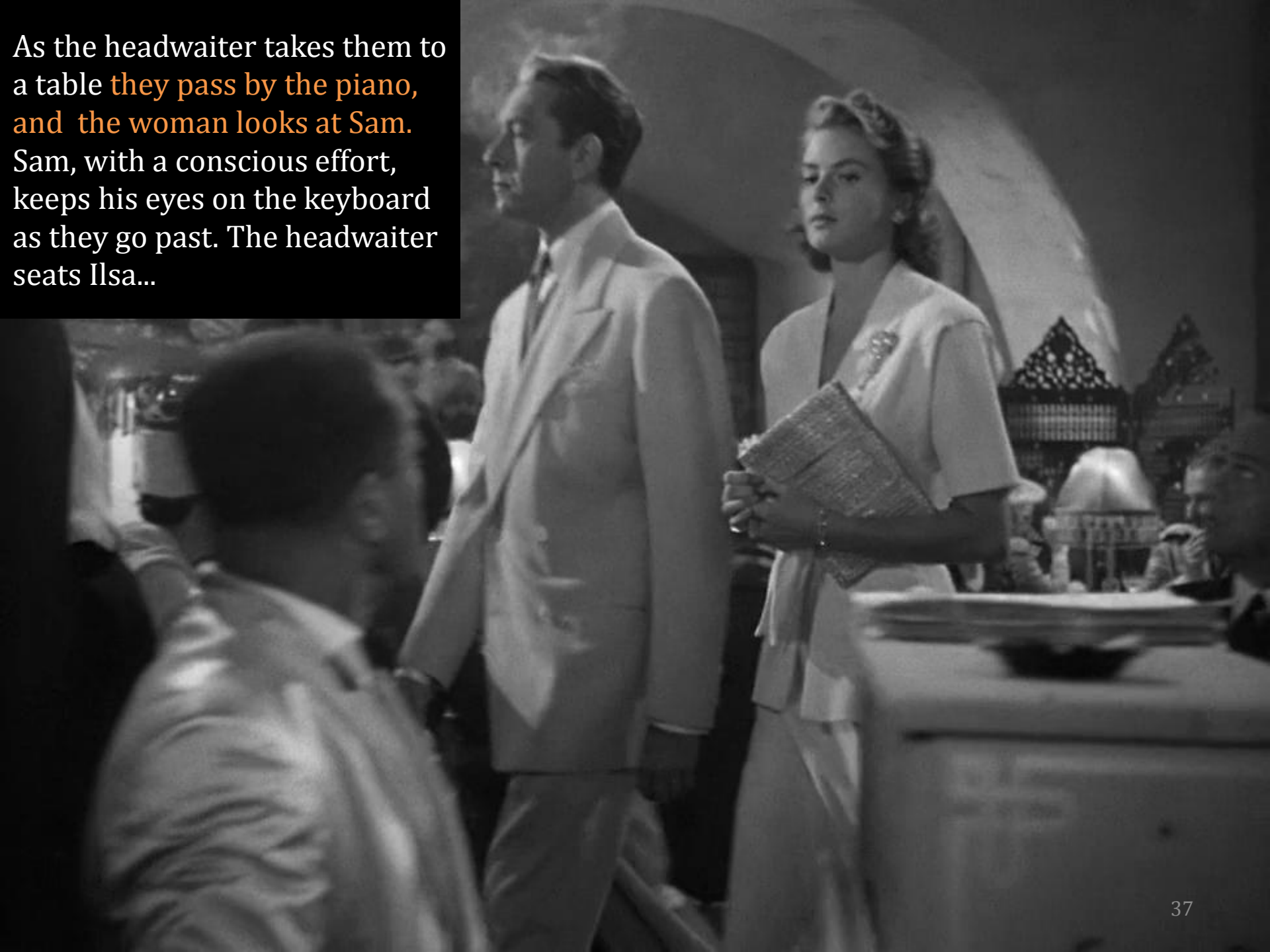
  HMDB, Olympic Sports,
  UCF50, UCF101, …

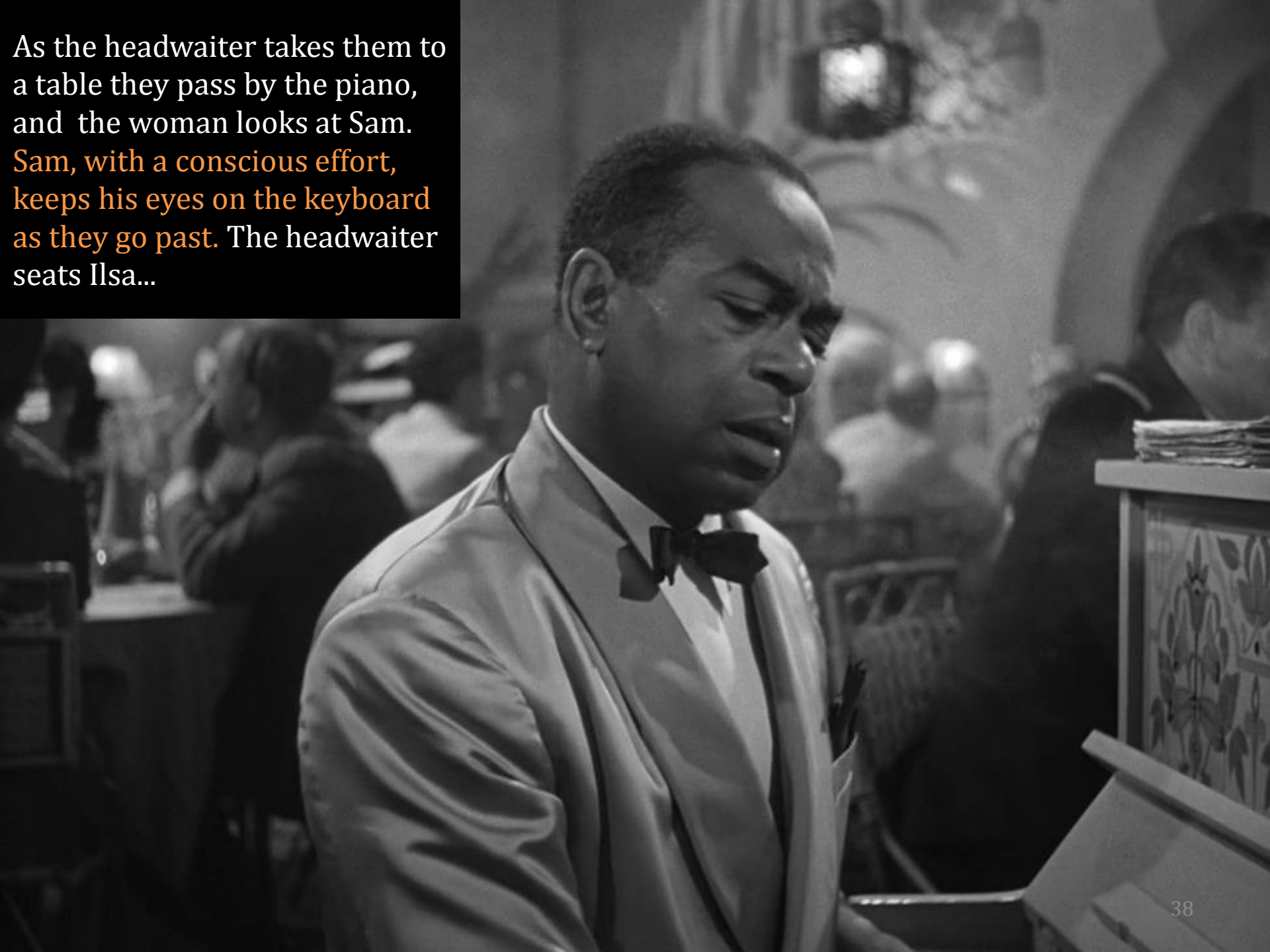
  ⇨ - Very time-consuming



- 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

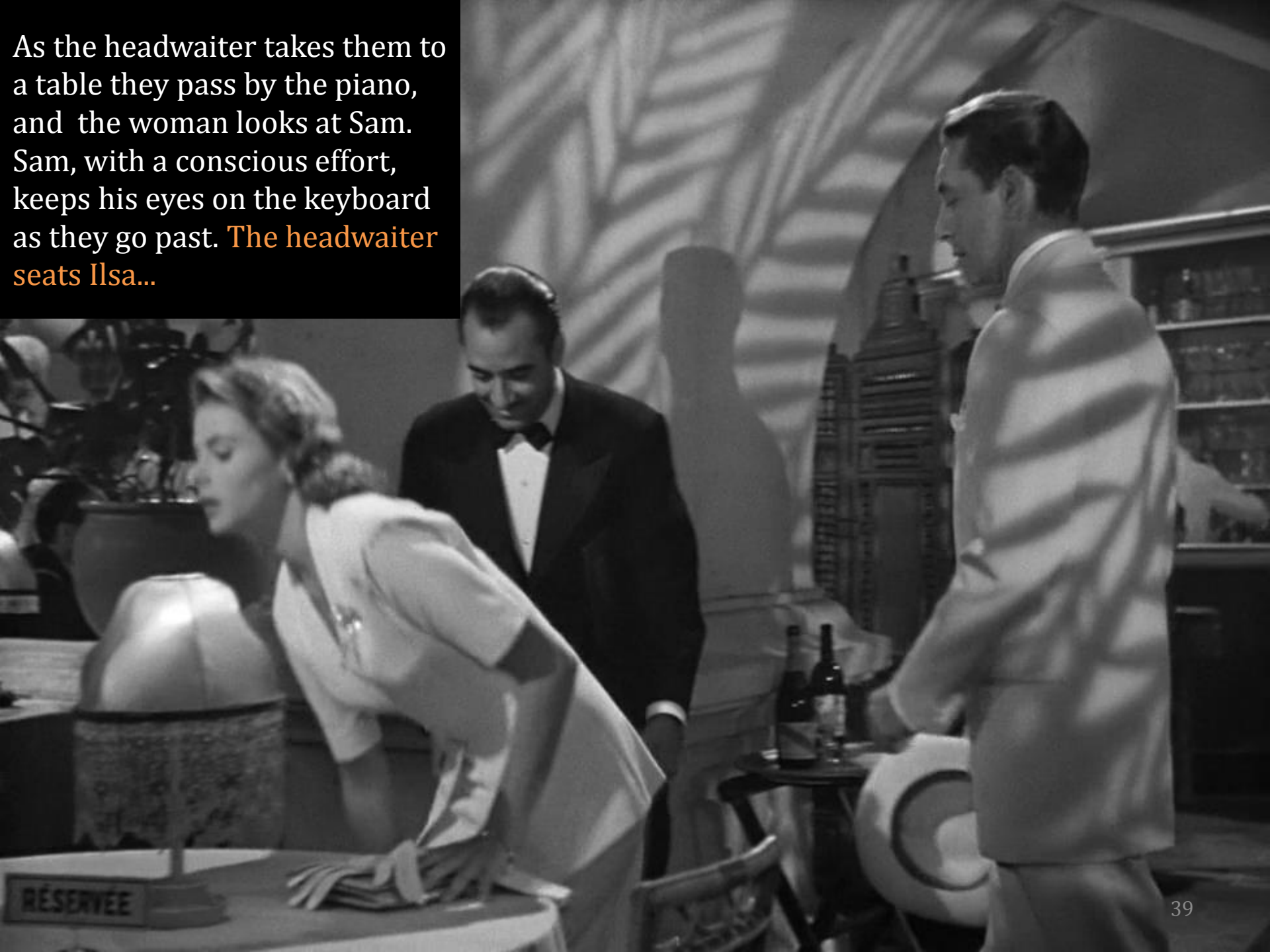

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...


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...



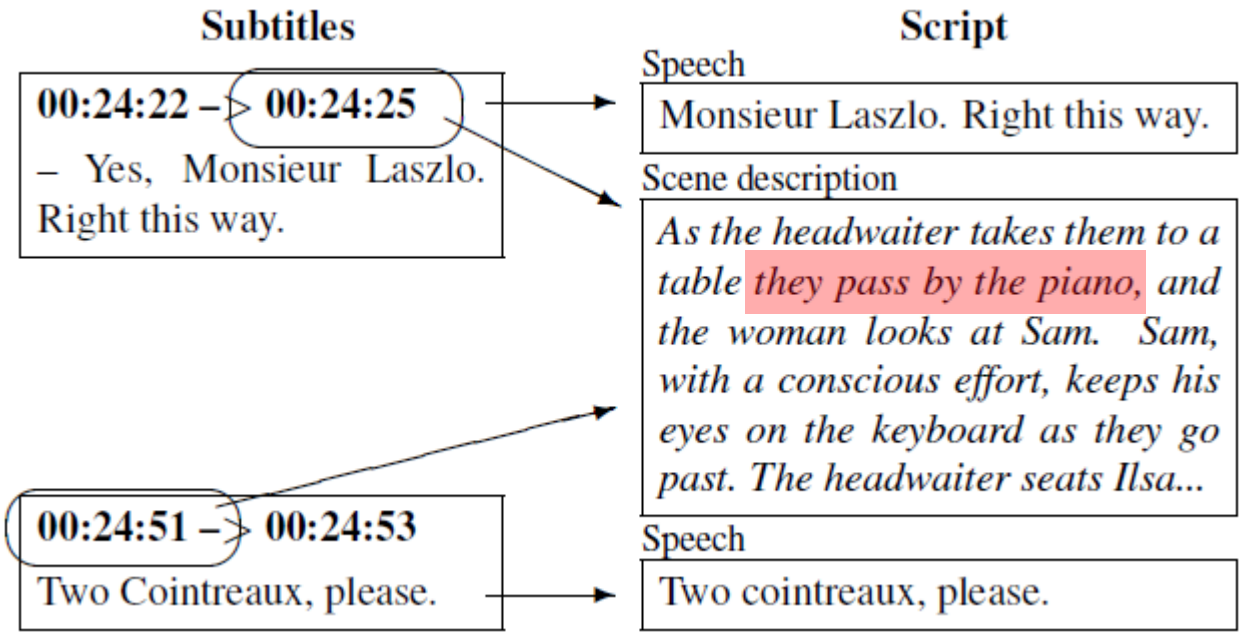
Scripts as weak supervision

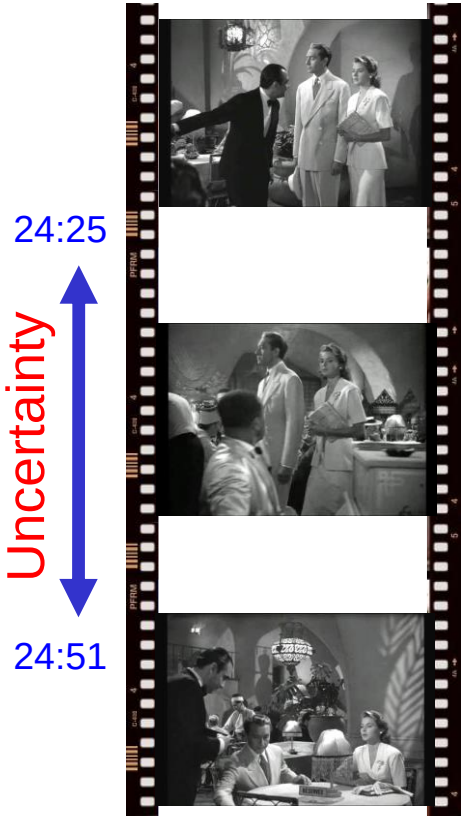
Challenges:

- 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

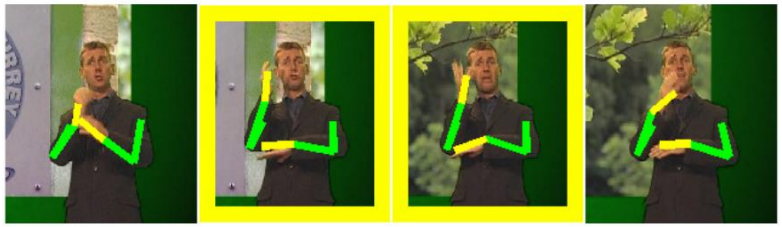




Previous work

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

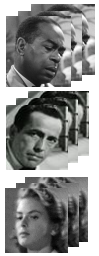
$$\frac{1}{N}\|Z - \phi(X)w - b\|_F^2 + \lambda_1 \, Tr(w^T \, w)$$

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

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

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

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

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 :

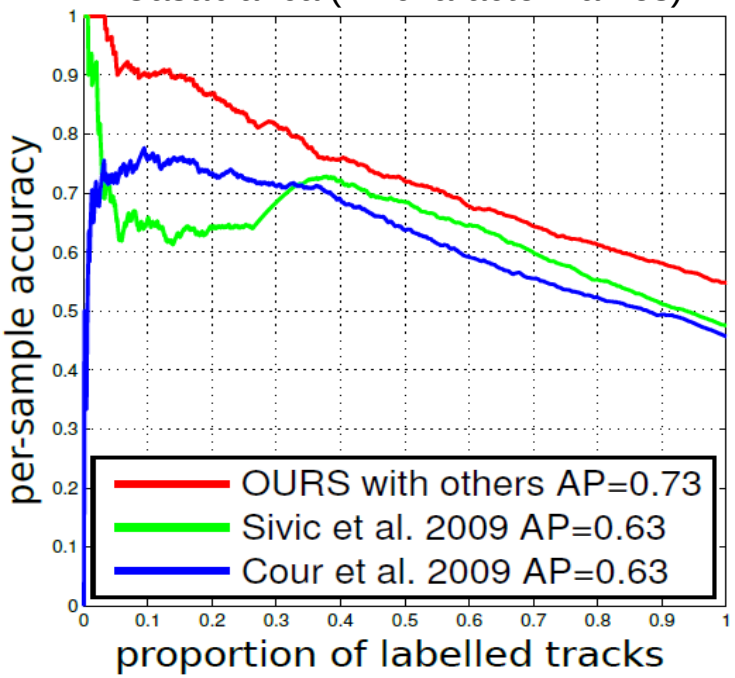
$$\sum_{n \in \mathcal{N}_i} t_{na} \geq 1$$

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

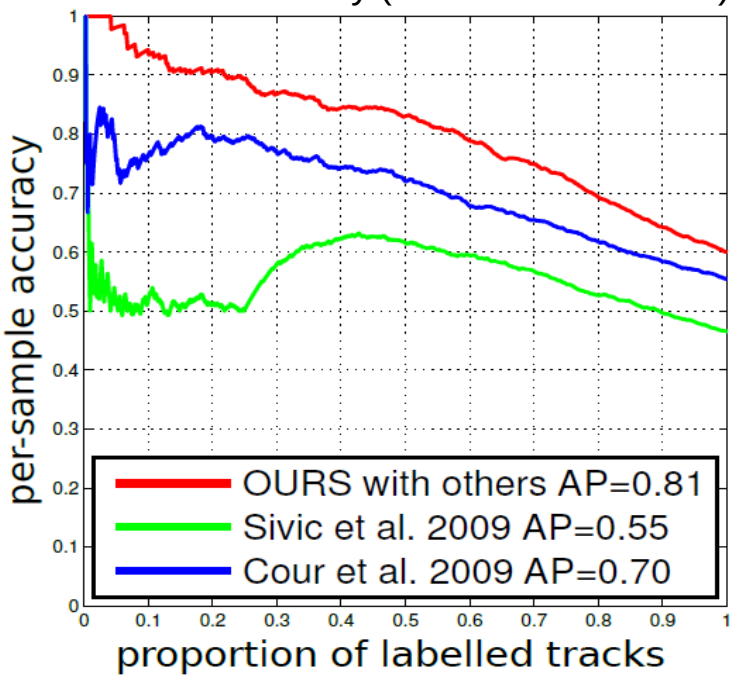
$$\sum_{n \in \mathcal{N}_i} z_{np} \, t_{na} \geq 1$$

Results for Person Labelling

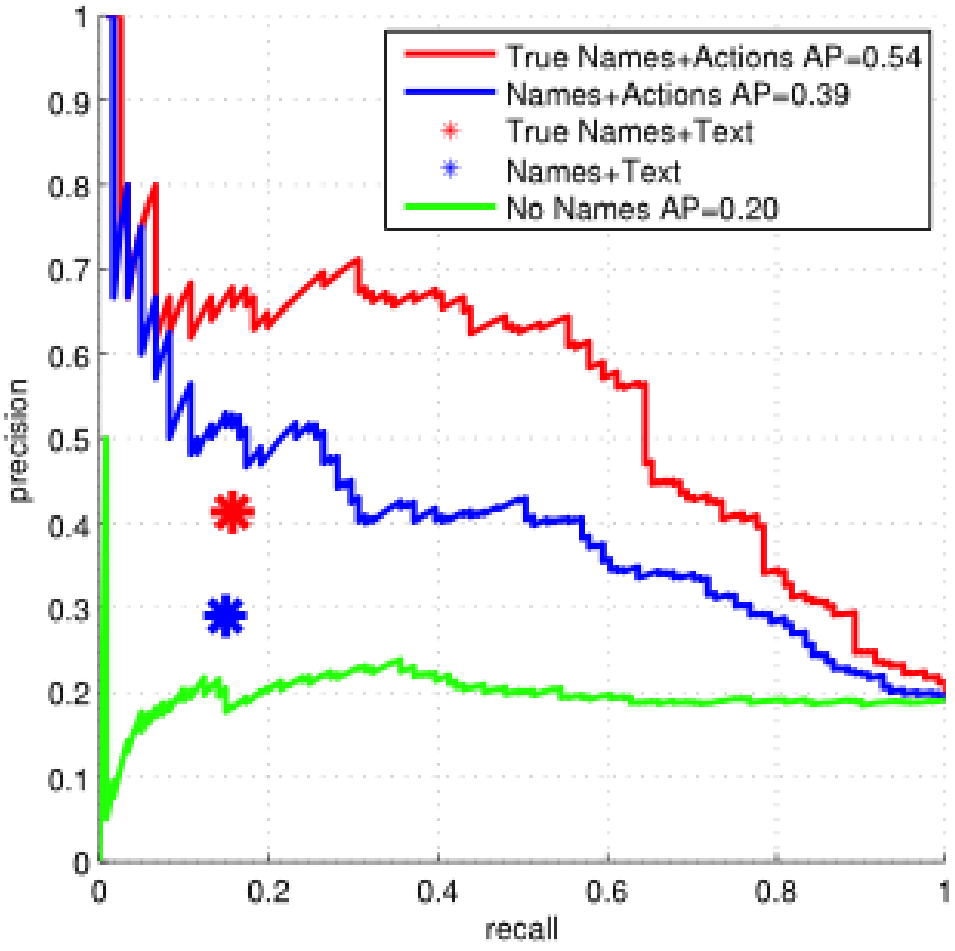
Casablanca (17 character names)



American beauty (11 character names)

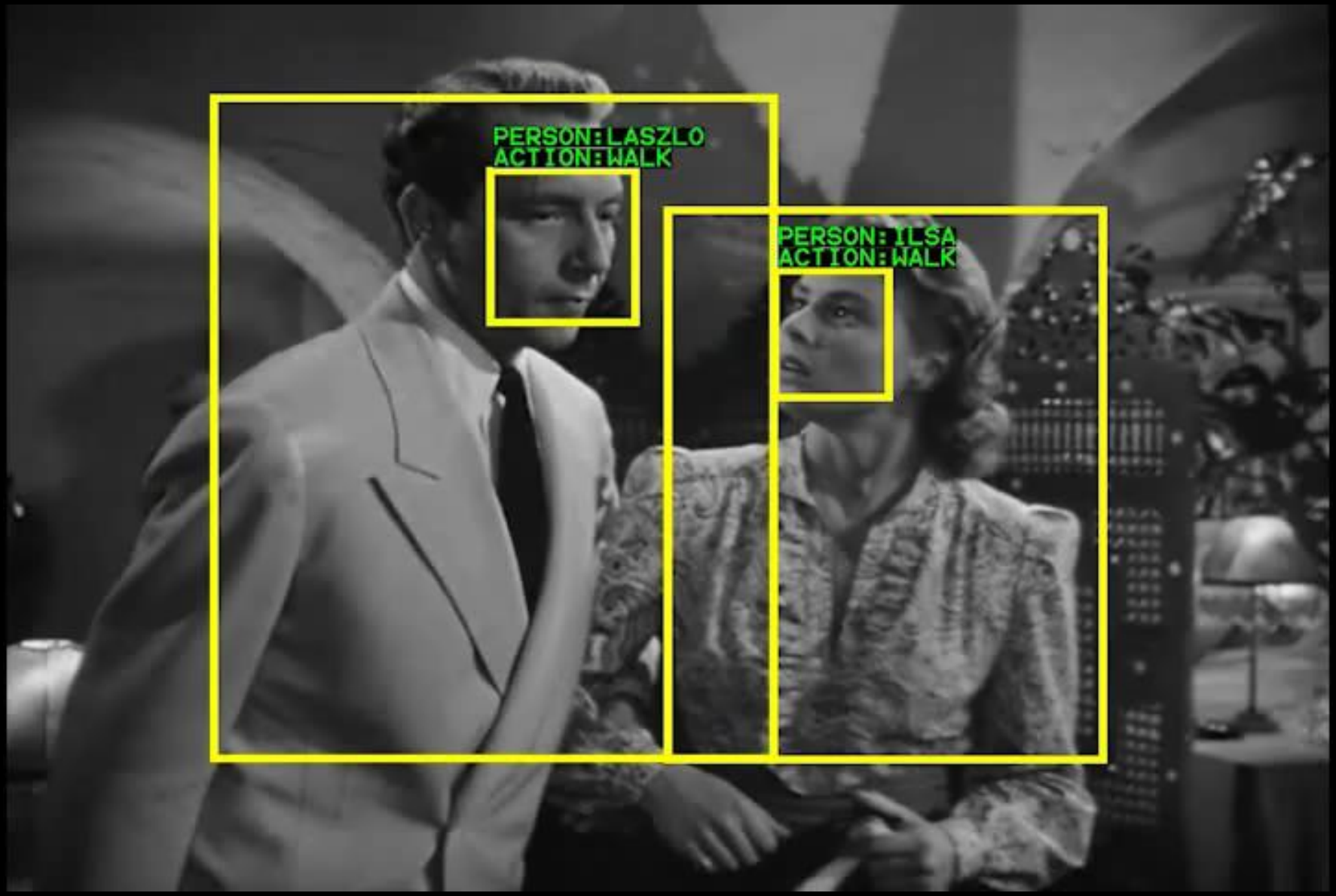


Results for Person + Action Labelling



Casablanca, Walking

Finding Actions and Actors in Movies



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

References to *some* related work

Video and language

- Laptev et al. Learning realistic human actions from movies, In CVPR 2008.
- Duchenne et al., Automatic Annotation of Human Actions in Video, ICCV 2009.
- Buehler et al., Learning sign language by watching TV (using weakly aligned subtitles), CVPR 2009.
- Gupta et al., Understanding Videos, Constructing Plots: Learning a Visually Grounded Storyline Model from Annotated Videos, CVPR 2009.
- Das et al., A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching, CVPR'13
- Ramanathan et al., Video Event Understanding using Natural Language Descriptions, ICCV 2013.
- Bojanowski et al., Finding Actors and Actions in Movies, ICCV 2013
- Cynthia, et al., Learning to parse natural language commands to a robot control system, Experimental Robotics, 2013.
- Ramanathanet al., Joint person naming in videos and coreference resolution in text, ECCV 2014
- Bojanowski et al., Weakly supervised action labeling in videos under ordering constraints, ECCV 2014

What are the challenges?

- **Large variations in appearance:** occlusions, non-rigid motion, view-point changes, clothing…

Action *Hugging*:



…

- **Manual collection of training samples is prohibitive:** many action classes, rare occurrence





…

- **Action vocabulary is not well-defined**

Action *Open*:




…

What are the challenges?

- **Large variations in appearance:** occlusions, non-rigid motion, view-point changes, clothing…

Action *Hugging*:


…

- **Manual collection of training samples is prohibitive:** many action classes, rare occurrence




…

- **Action vocabulary is not well-defined**

Action *Open*:



…

Where is computer vision going next?

Is image/video classification the right problem?

- Is action vocabulary well-defined?

Examples of “Open” action:



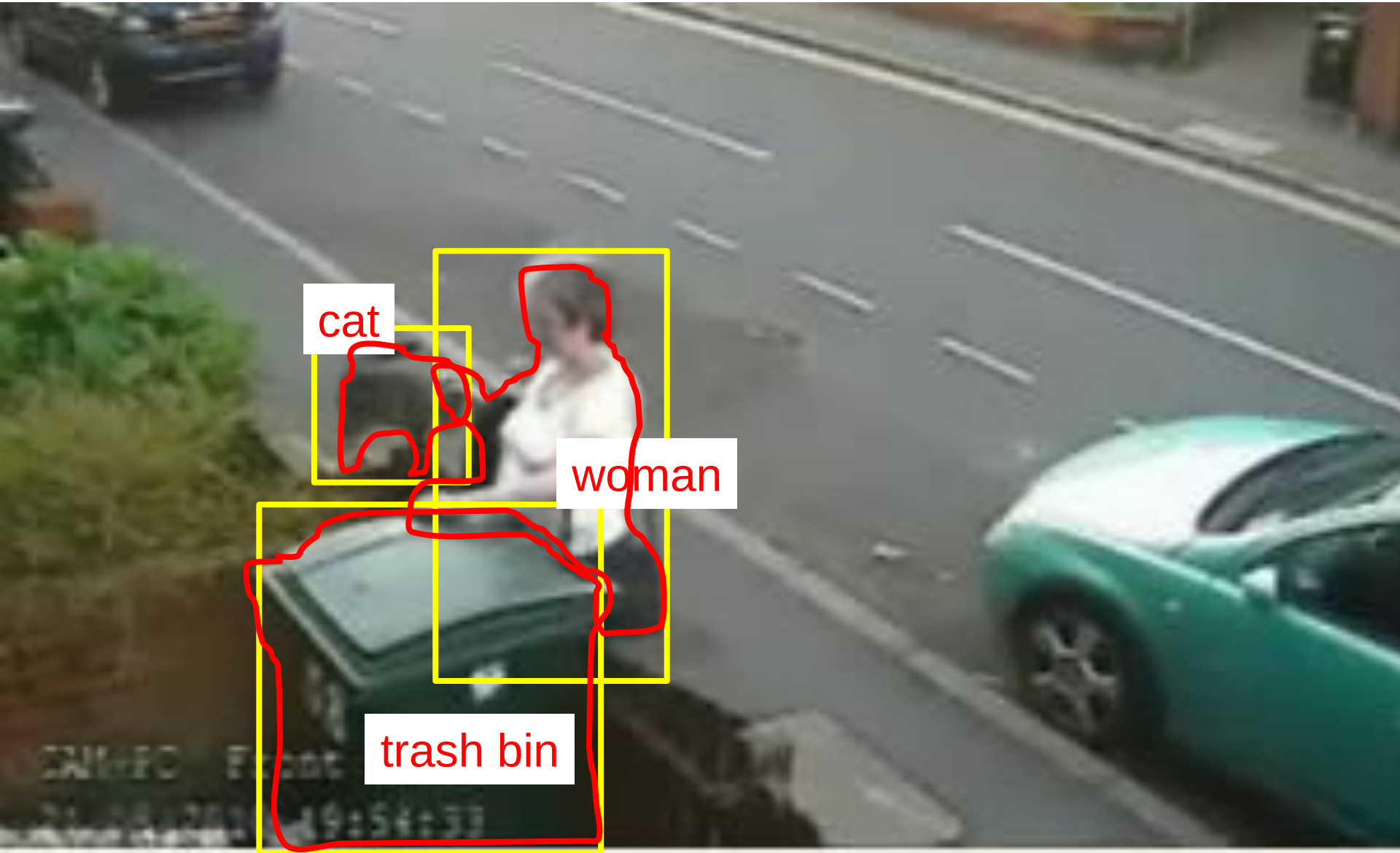



- What granularity of action vocabulary shall we consider?

Source: http://www.youtube.com/watch?v=eYdUZdan5i8

Do we want to learn *person-throws-cat-into-trash-bin* classifier?

Can object recognition help?



Limitations of Current Methods

Limitations of Current Methods

What is unusual in this scene?

Is this scene dangerous?

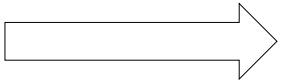
What is intention of this person?

What is unusual in this scene?

Next challenge

Shift the focus of computer vision

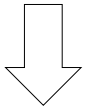
Object, scene and action recognition



Recognition of objects' function and people's intentions

Is this a picture of a dog?
Is the person running in this video?

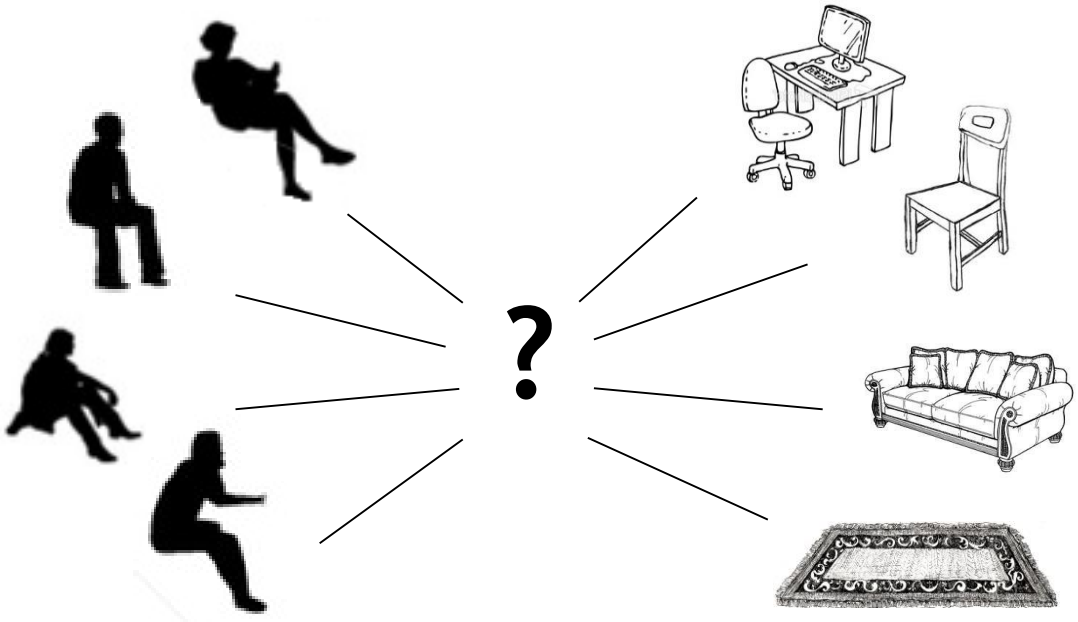
What people do with objects?
How they do it?
For what purpose?



Enable new applications

Motivation

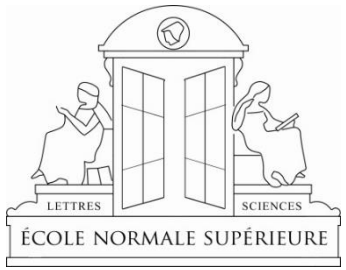
- Exploit the link between human pose, action and object function.



- Use human actors as active sensors to reason about the surrounding scene.

Scene semantics from long-term observation of people

V. Delaitre, D. F. Fouhey, I. Laptev,
J. Sivic, A. Gupta, A. Efros



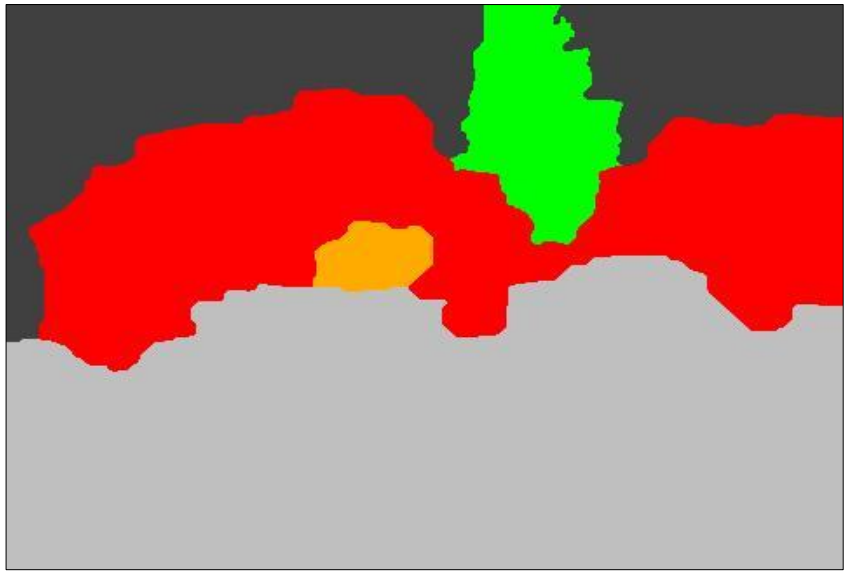
Goal

Recognize objects by the way people interact with them.

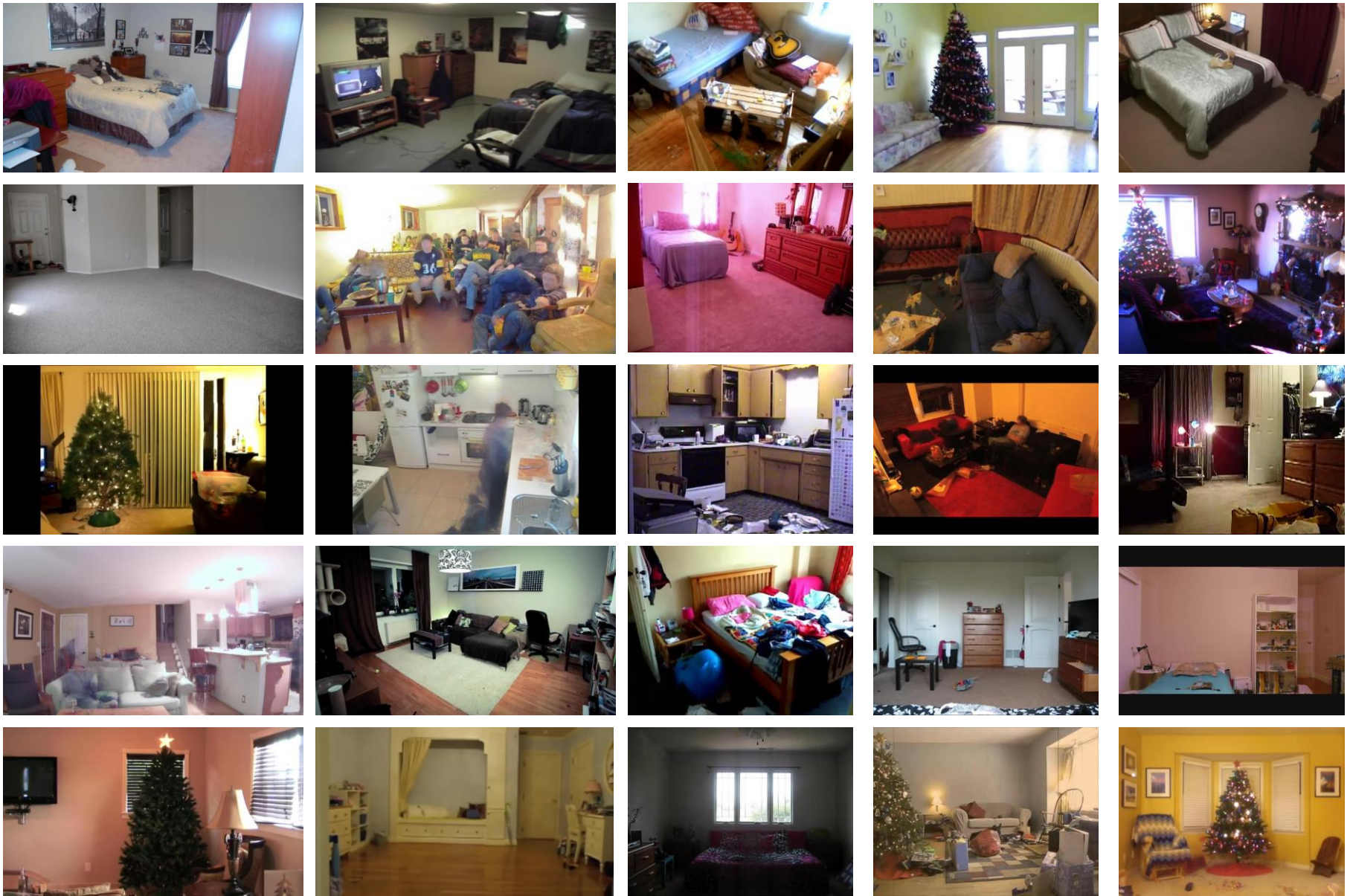
Time-lapse “Party & Cleaning” videos

Lots of person-object interactions, many scenes on YouTube

Semantic object segmentation



New “Party & Cleaning” dataset



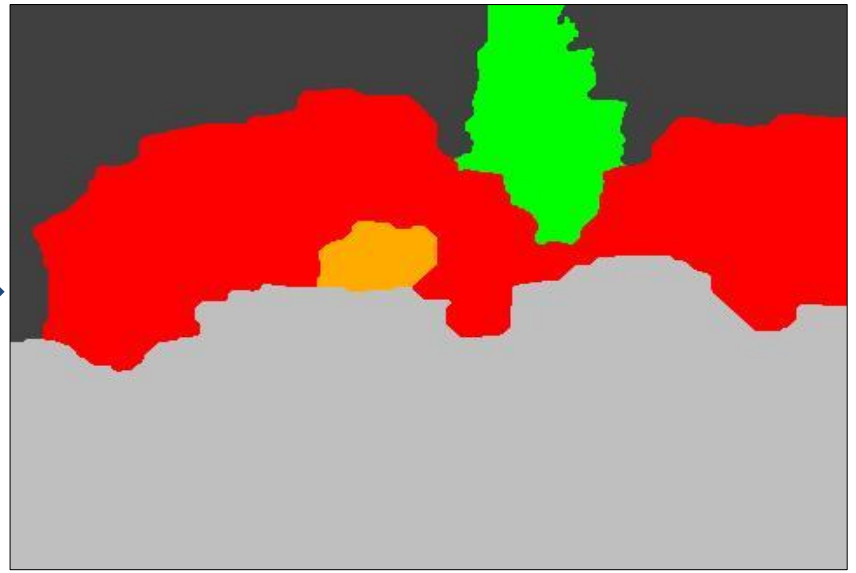
Goal

Recognize objects by the way people interact with them.

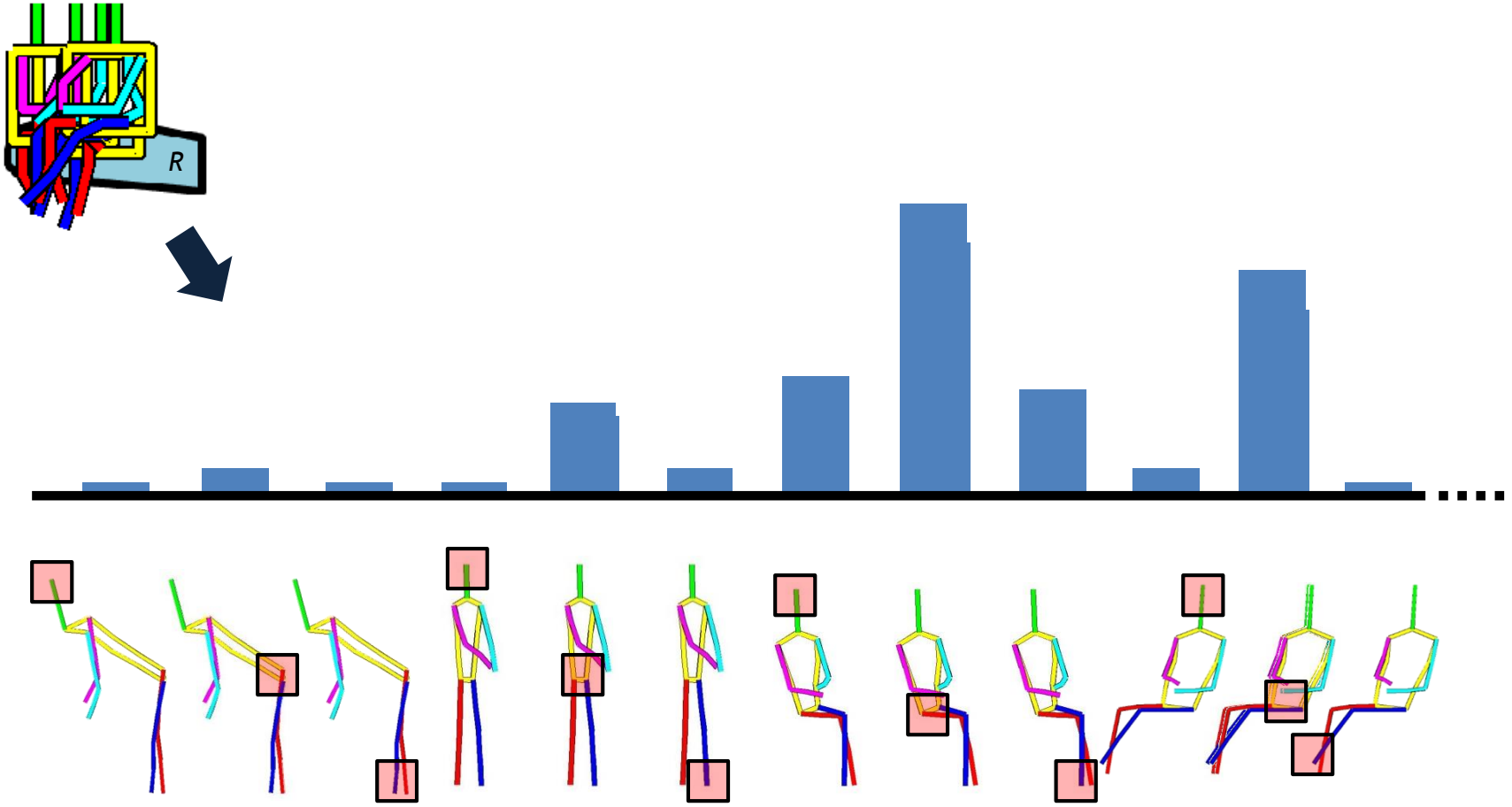
Time-lapse “Party & Cleaning” videos

Lots of person-object interactions, many scenes on YouTube

Semantic object segmentation



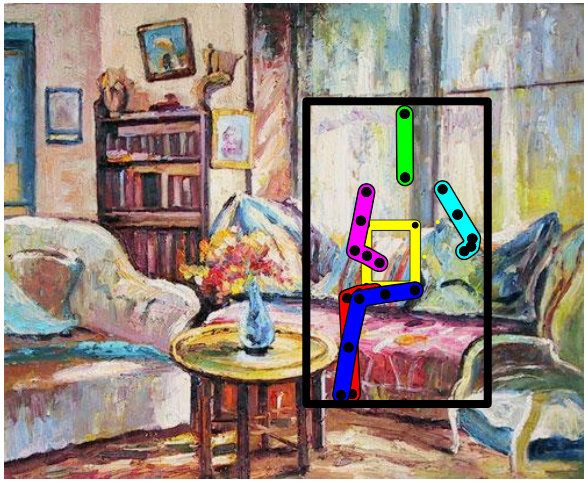
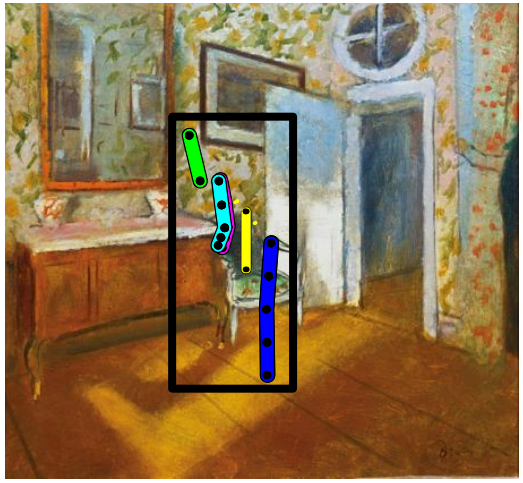
Pose histogram

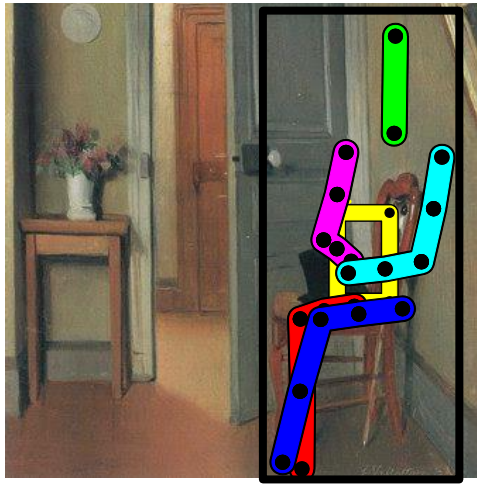
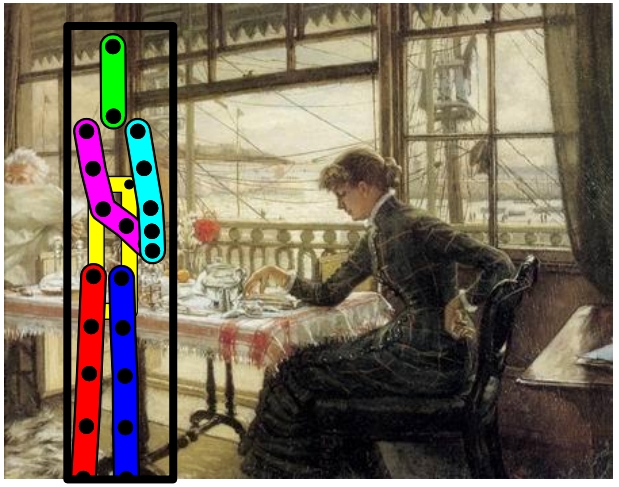
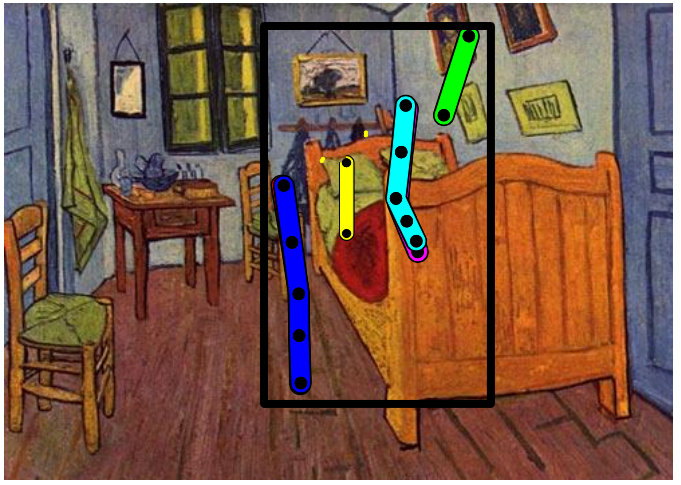


Some qualitative results

Using our model as pose prior

Given a bounding box and the ground truth segmentation, we fit the pose clusters in the box and score them by summing the joint's weight of the underlying objects.

Input image