INRIA RENNES

Action class detection and recognition in realistic video

[ICCV07]

Learning realistic human actions from movies

[CVPR08]

Ivan Laptev, Patrick Pérez
Marcin Marszalek, Cordelia Schmid
Benjamin Rozenfeld

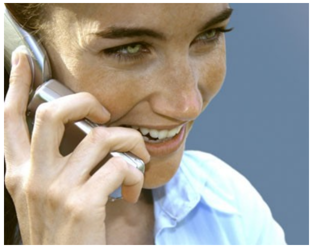
INRIA Rennes, France
INRIA Grenoble, France
Bar-Ilan University, Israel

Presenter: Scott Satkin

Slide Courtesy: Ivan Laptev

Human actions: Motivation

- Huge amount of video is available and growing
- Human actions are major events in movies, TV news, personal video …


Action recognition useful for:

- Content-based browsing
  e.g. fast-forward to the next goal scoring scene
- Video recycling
  e.g. find "Bush shaking hands with Putin"
- Human scientists
  influence of smoking in movies on adolescent smoking

What are human actions?

Definition 1:

- **Physical body motion**

  [Niebles et al.'06, Shechtman&Irani'05,
  Dollar et al.'05, Schuldt et al.'04, Efros et al.'03
  Zelnik-Manor&Irani'01, Yacoob&Black'98,
  Polana&Nelson'97, Bobick&Wilson'95, …]

KTH action dataset

Definition 2:

- **Interaction with environment on specific purpose**

  same physical motion -- different actions depending on the context

Context defines actions

Challenges in action recognition

- **Similar problems to static object recognition:**
  variations in views, lightning, background, appearance, …
- **Additional problems:** *variations in individual motion; camera motion*

Example:

Drinking

Smoking

Difference in shape

Difference in motion

Both actions are similar in overall shape (human posture) and motion (hand motion)

Data variation for actions might be higher than for objects

But: *Motion provides an additional discriminative cue*

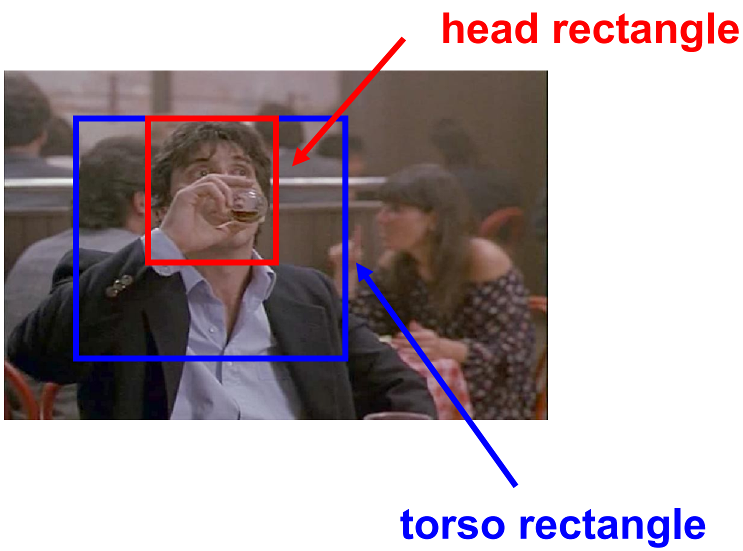
Action dataset and annotation

- No datasets with realistic action classes are available
- This work: *first attempt to approach action detection and recognition in real movies*: “Coffee and Cigarettes”; “Sea of Love”

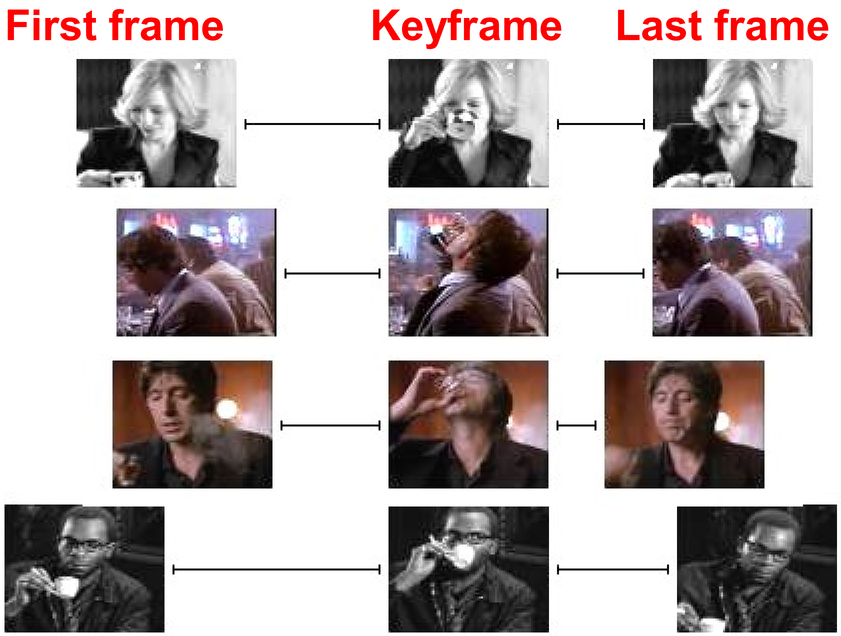
“*Drinking*”: 159 annotated samples

“*Smoking*”: 149 annotated samples

Spatial annotation



Temporal annotation



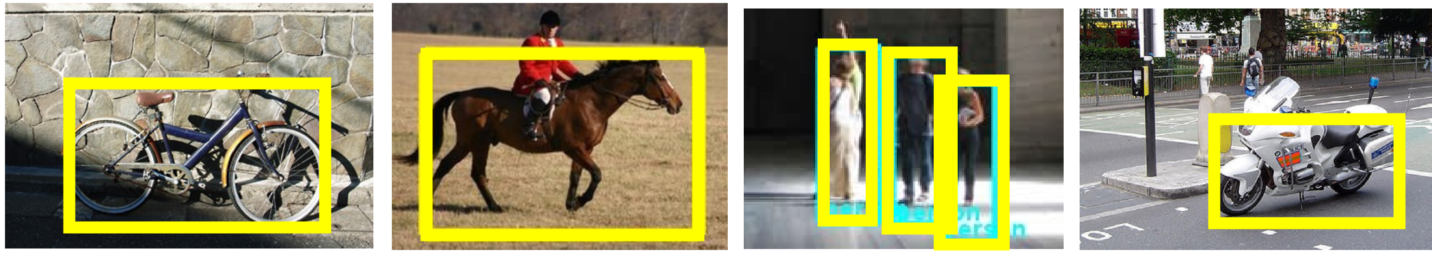
“Drinking” action samples

training samples

test samples

Actions == space-time objects?

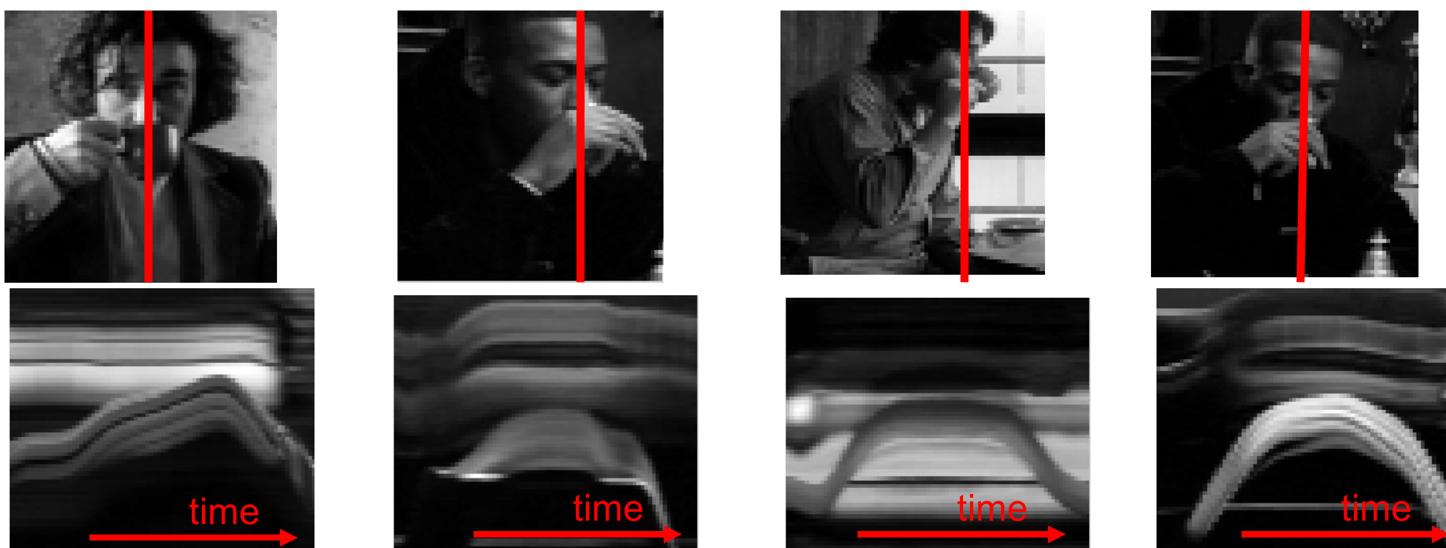
“stable-view” objects



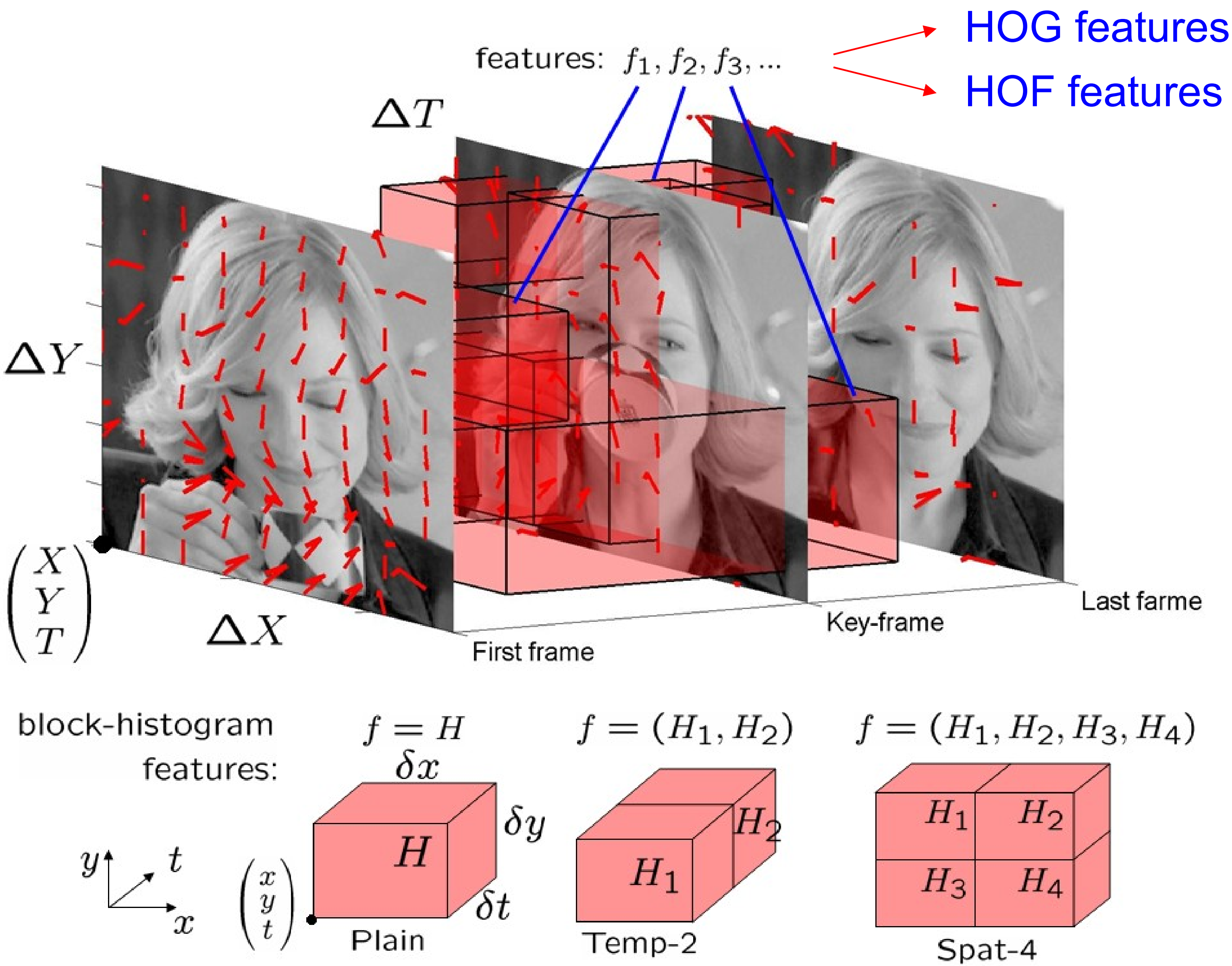
“atomic” actions

car exit phoning smoking hand shaking drinking

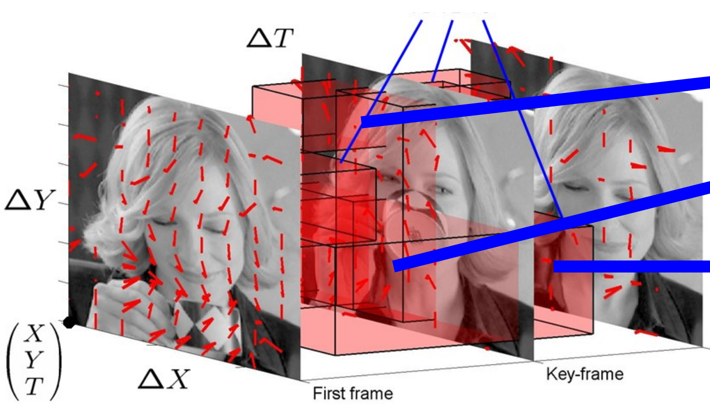
Objective: take advantage of space-time shape



Action features

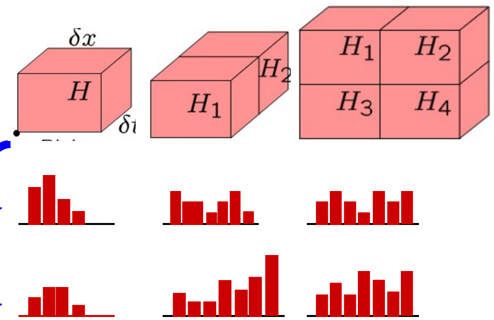


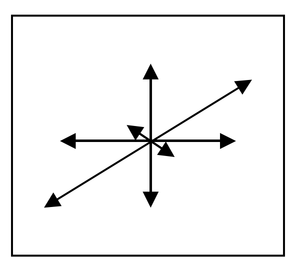
Histogram features



~10^7 cuboid features
Choosing 10^3 randomly

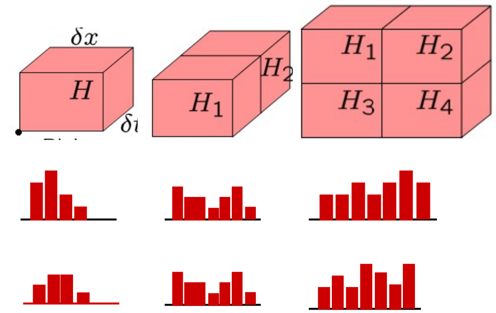
HOG: histograms of
oriented gradient

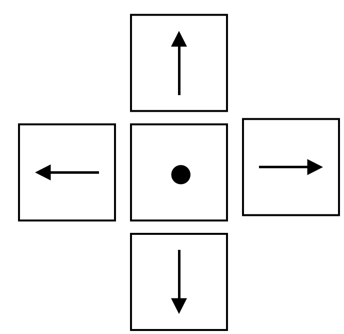




4 grad. orientation bins

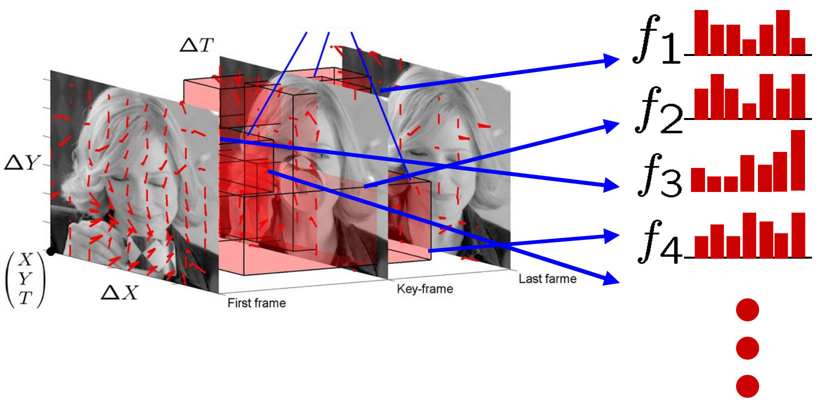
HOF: histograms of
optic flow





4 OF direction bins
+ 1 bin for no motion

Action learning



boosting

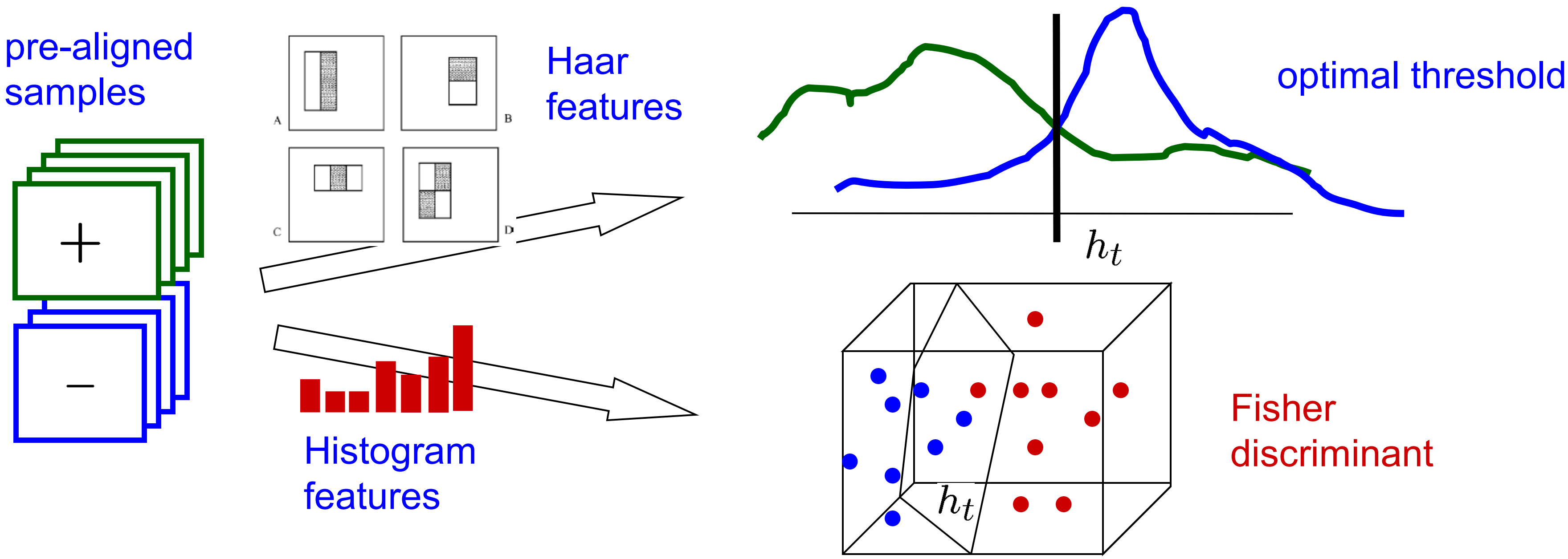
selected features

$$H(z) = \mathrm{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(f_t)\right)$$

weak classifier

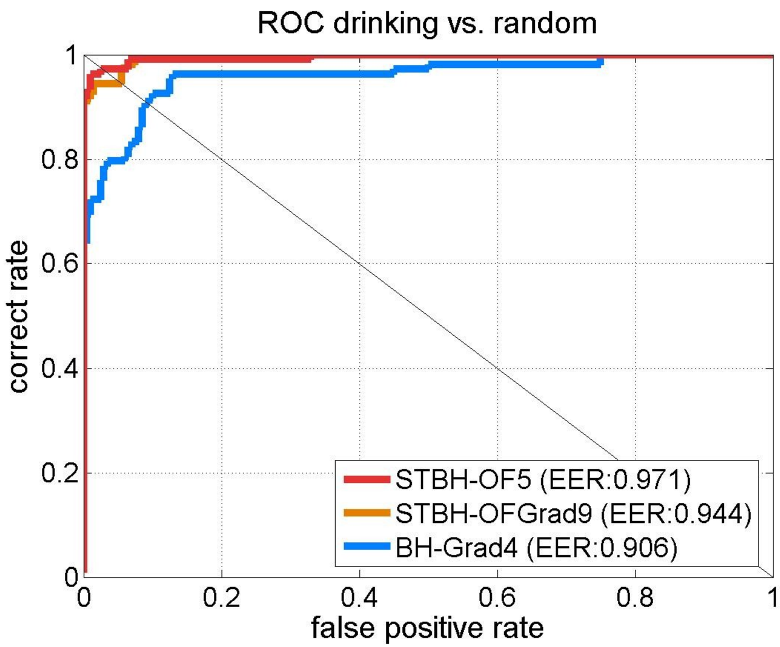
AdaBoost:

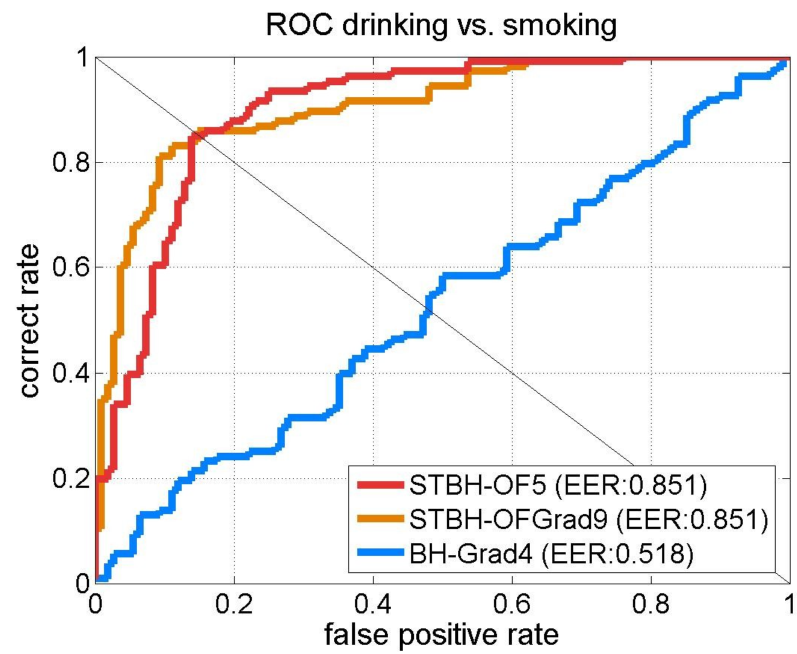
- Efficient discriminative classifier [Freund&Schapire'97]
- Good performance for face detection [Viola&Jones'01]



Action classification test

Random motion patterns





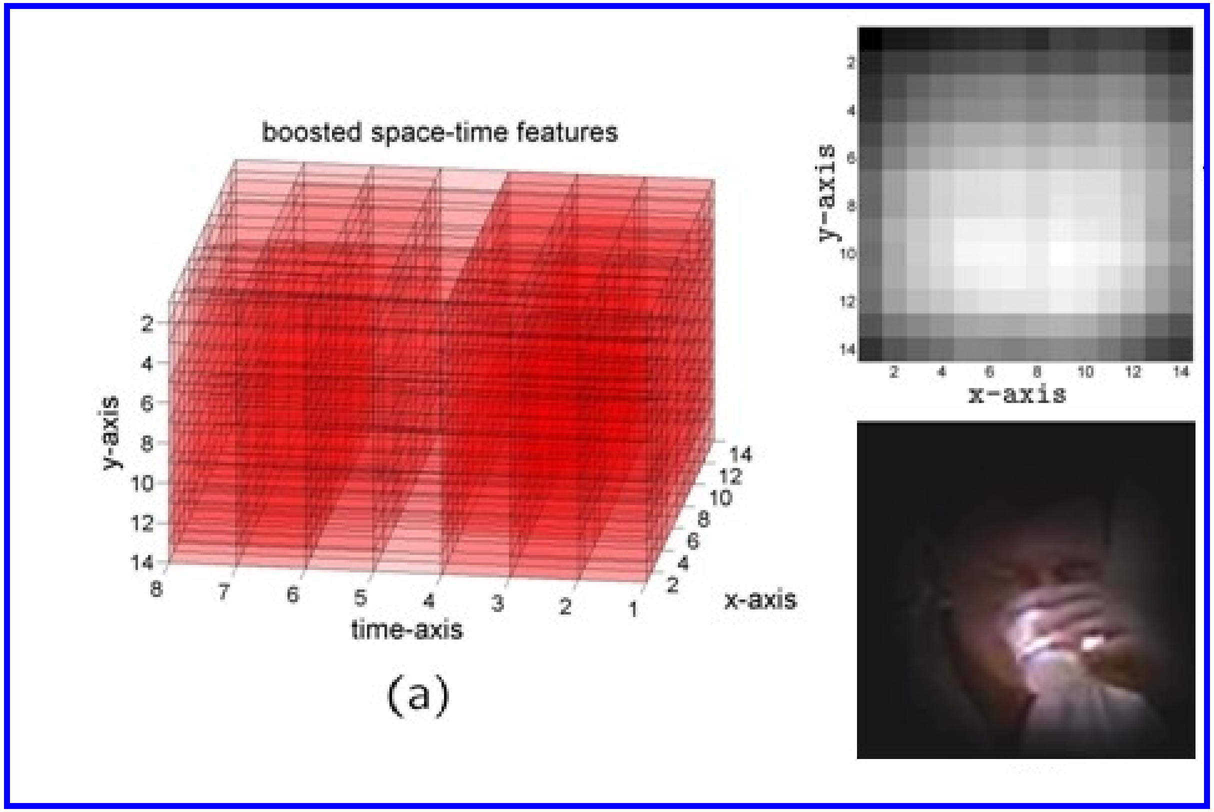
- Additional shape information does not seem to improve the space-time classifier
- Space-time classifier and static key-frame classifier might have complementary properties

Classifier properties

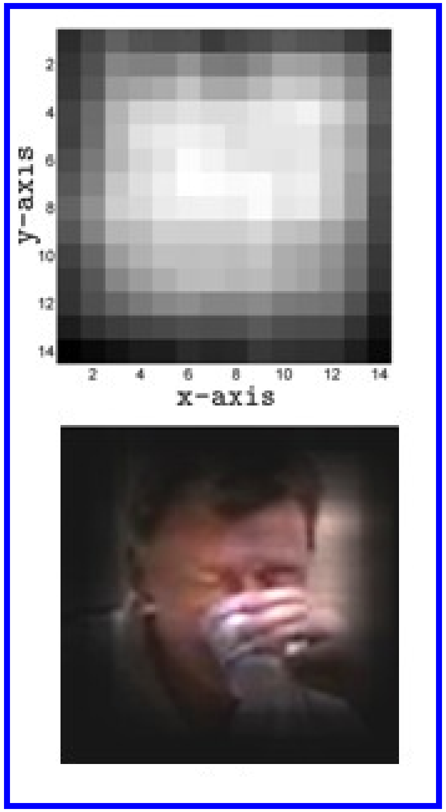
Compare selected features by

- Space-time action classifier (HOF features)
- Static key-frame classifier (HOG features)

Training output: Accumulated feature maps



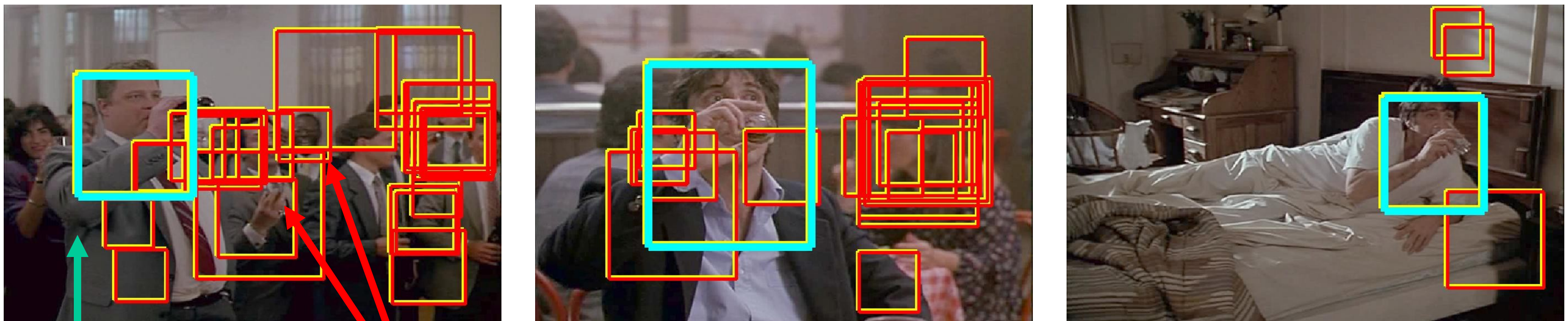
Space-time classifier



Static keyframe classifier

Keyframe priming

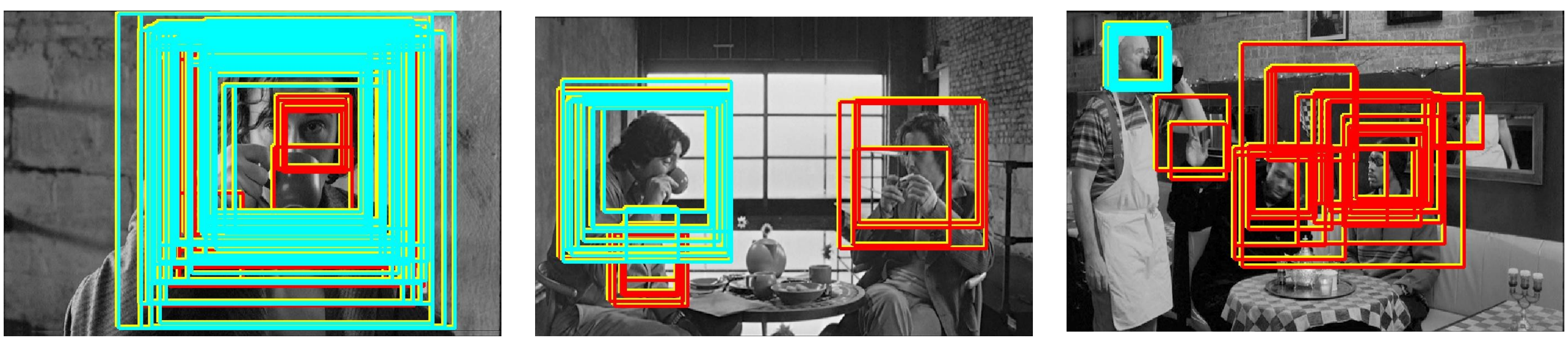
Training



Positive training sample

Negative training samples

Test



Action detection

Test set:

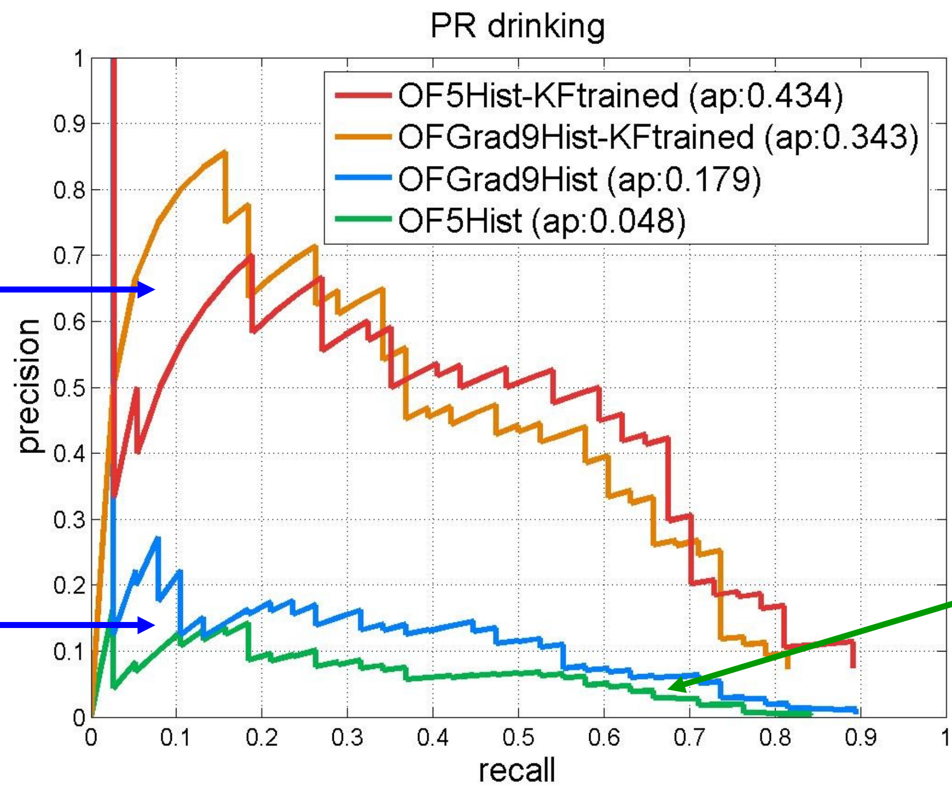
- 25min from “Coffee and Cigarettes” with GT 38 drinking actions
- No overlap with the training set in subjects or scenes

Detection:

- search over all space-time locations and spatio-temporal extents

Keyframe priming

No Keyframe priming



Similar approach to Ke, Sukthankar and Hebert, ICCV05

Test episode

Summary

- First attempt to address human action in real movies
- Action detection/recognition seems possible under hard realistic conditions (variations across views, subjects, scenes, etc…)
- Separate learning of shape/motion information results in a large improvement

Future

- Need realistic data for 100’s of action classes
- Explicit handling of actions under multiple views
- Combining action classification with text

Access to realistic human actions

Web *video* search

- Useful for some action classes: *kissing, hand shaking*
- Very noisy or not useful for the majority of other action classes
- Examples are frequently non-representative

Goodle Video, YouTube, MyspaceTV, …

Access to realistic human actions

Web *video* search

- Useful for some action classes: *kissing, hand shaking*
- Very noisy or not useful for the majority of other action classes
- Examples are frequently non-representative

Goodle Video, YouTube, MyspaceTV, …

Actions in movies

- Realistic variation of human actions
- Many classes and many examples per class



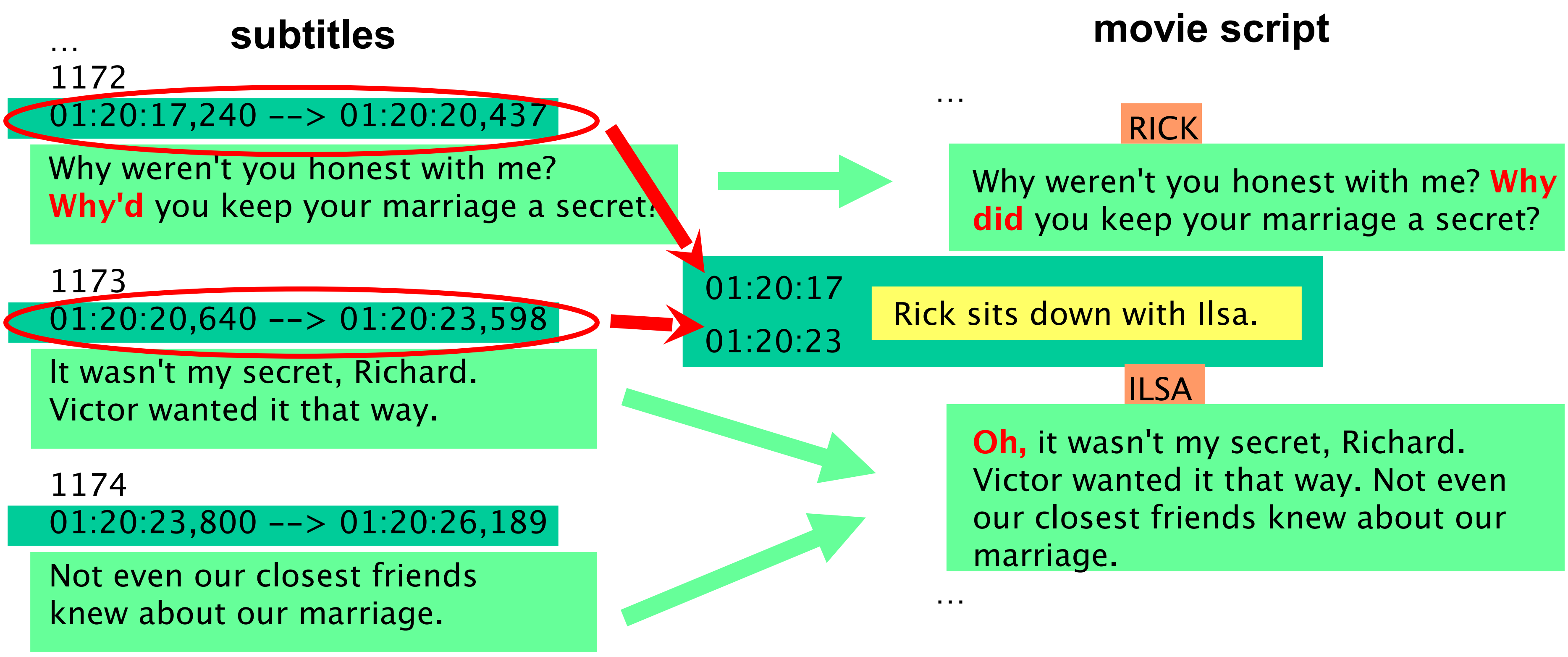



Problems:

- Typically only a few class-samples per movie
- Manual annotation is very time consuming

Automatic video annotation using scripts [Everingham et al. BMVC06]

- Scripts available for >500 movies (no time synchronization)
  www.dailyscript.com, www.movie-page.com, www.weeklyscript.com …
- Subtitles (with time info.) are available for the most of movies
- Can transfer time to scripts by text alignment



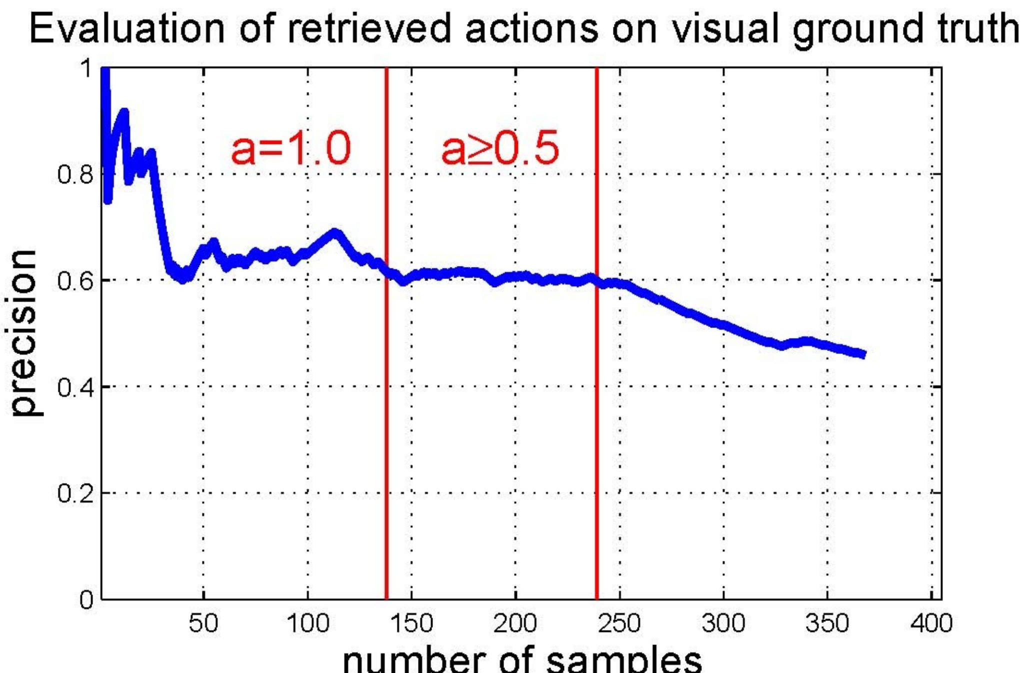
Script-based action annotation

- **On the good side:**
  - Realistic variation of actions: subjects, views, etc…
  - Many examples per class, many classes
  - No extra overhead for new classes
  - Actions, objects, scenes and their combinations
  - Character names may be used to resolve “who is doing what?”

- **Problems:**
  - No spatial localization
  - Temporal localization may be poor
  - Missing actions: e.g. scripts do not always follow the movie
  - Annotation is incomplete, not suitable as ground truth for testing action detection
  - Large within-class variability of action classes *in text*

Script alignment: Evaluation

- Annotate action samples *in text*
- Do automatic script-to-video alignment
- Check the correspondence of actions in scripts and movies



a: quality of subtitle-script matching

Example of a "visual false positive"

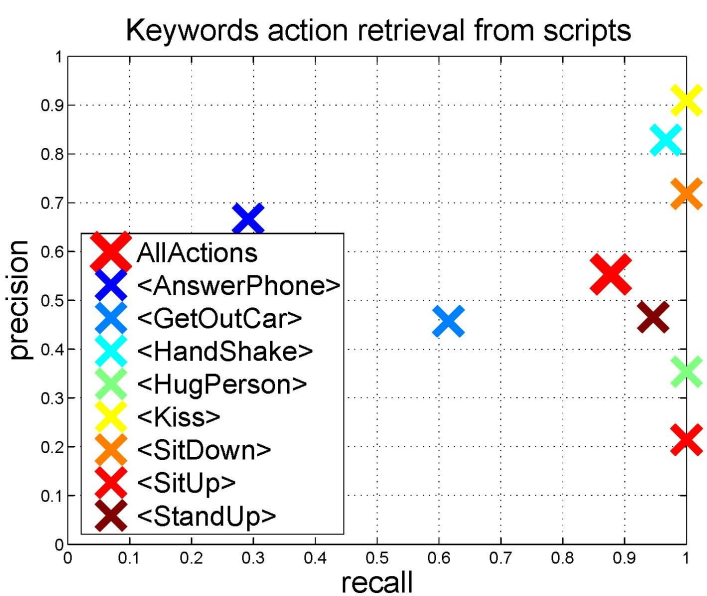
A black car pulls up, two army officers get out.

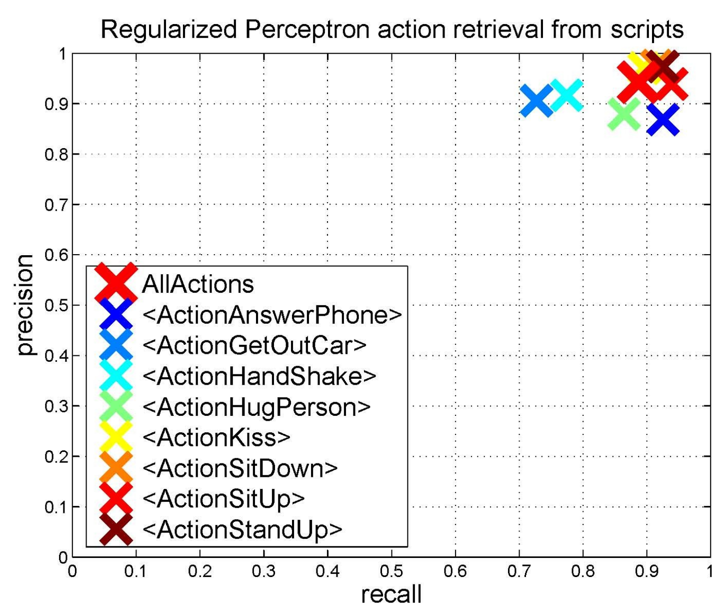
Text-based action retrieval

- Large variation of action expressions in text:

GetOutCar action:	*"… Will gets out of the Chevrolet. …"* *"… Erin exits her new truck…"*
Potential false positives:	*"…About to sit down, he freezes…"*

- => Supervised text classification approach





Movie actions dataset

12 movies

	<AnswerPhone>	<GetOutCar>	<HandShake>	<HugPerson>	<Kiss>	<SitDown>	<SitUp>	<StandUp>	Total
False	5	6	9	7	10	21	5	33	96
Correct	15	6	14	8	34	30	7	29	143
All	20	12	23	15	44	51	12	62	239
automatically labeled training set									
	22	13	20	22	49	47	11	48	232
manually labeled training set									
	23	13	19	22	51	30	10	49	217
test set									

20 different movies

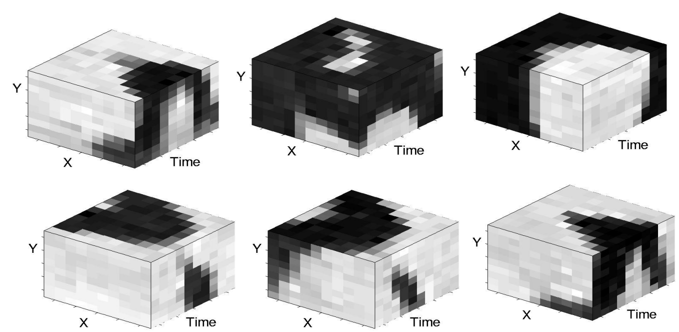
- Learn vision-based classifier from automatic training set
- Compare performance to the manual training set

Action Classification: Overview

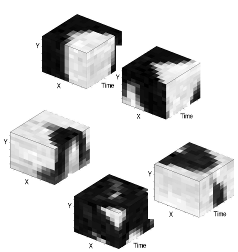
Bag of space-time features + multi-channel SVM

[Schuldt'04, Niebles'06, Zhang'07]

Collection of space-time patches

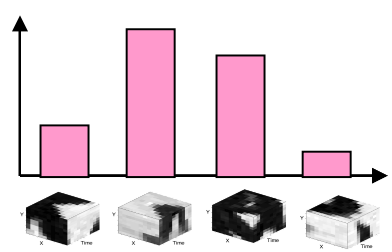


Visual vocabulary



HOG & HOF patch descriptors

Histogram of visual words



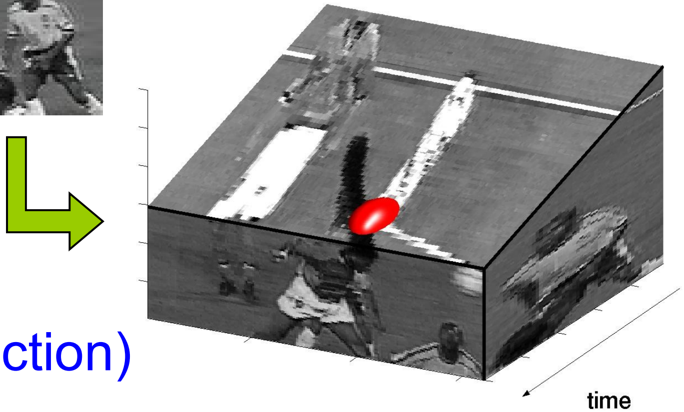
Multi-channel SVM Classifier

Space-Time Features: Detector

- Space-time corner detector
  [Laptev, IJCV 2005]

$$H = \det(\mu) + k\,\mathrm{tr}^3(\mu)$$

$$\mu = \begin{pmatrix} I_xI_x & I_xI_y & I_xI_t \\ I_xI_y & I_yI_y & I_yI_t \\ I_xI_t & I_yI_t & I_tI_t \end{pmatrix} * g(\cdot;\, \sigma, \tau)$$

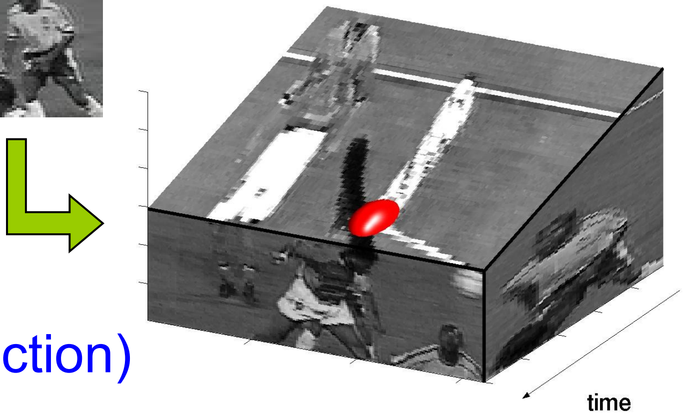


- Dense scale sampling (no explicit scale selection)

$$(\sigma^2, \tau^2) = \mathcal{S} \times \mathcal{T},\ \mathcal{S} = 2^{\{2,\ldots,6\}}, \mathcal{T} = 2^{\{1,2\}}$$

Space-Time Features: Detector

- Space-time corner detector
  [Laptev, IJCV 2005]



$$H = \det(\mu) + k\,\mathrm{tr}^3(\mu)$$

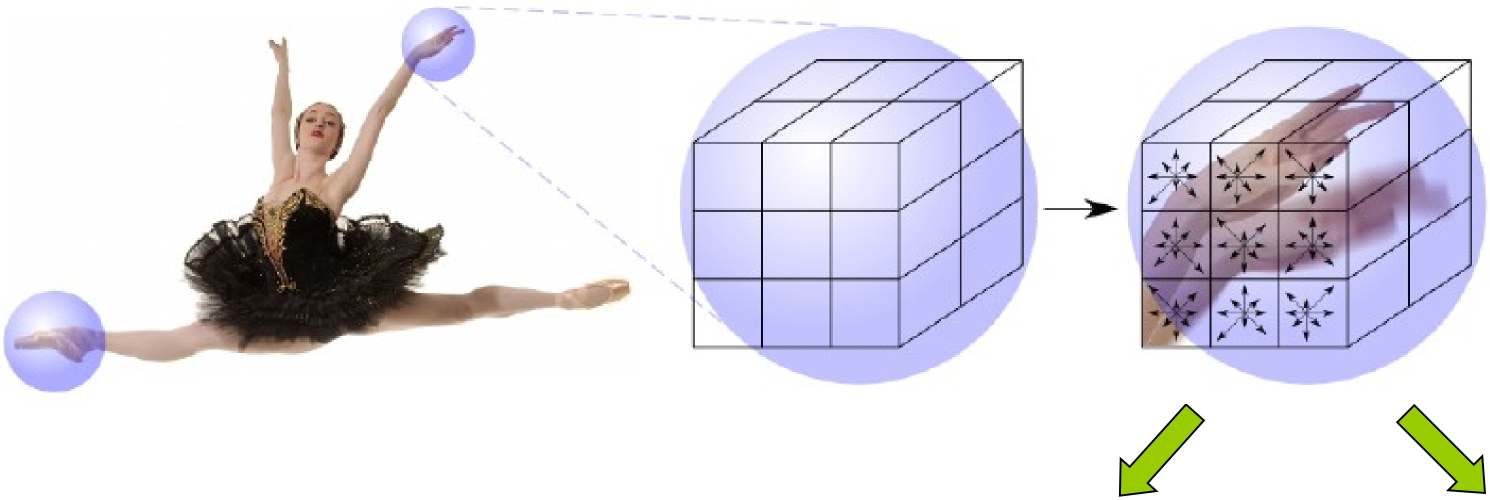
$$\mu = \begin{pmatrix} I_xI_x & I_xI_y & I_xI_t \\ I_xI_y & I_yI_y & I_yI_t \\ I_xI_t & I_yI_t & I_tI_t \end{pmatrix} * g(\cdot;\, \sigma, \tau)$$

- Dense scale sampling (no explicit scale selection)

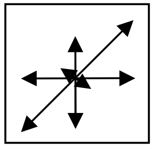
$$(\sigma^2, \tau^2) = \mathcal{S} \times \mathcal{T},\ \mathcal{S} = 2^{\{2,\ldots,6\}}, \mathcal{T} = 2^{\{1,2\}}$$

Space-Time Features: Descriptor

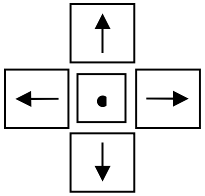
Multi-scale space-time patches
from corner detector

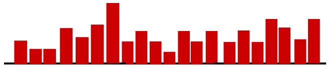


Histogram of
oriented spatial
grad. (HOG)

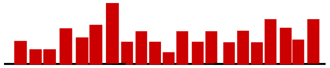


Histogram
of optical
flow (HOF)





3x3x2x4bins **HOG**
descriptor



3x3x2x5bins **HOF**
descriptor

Public code available at
www.irisa.fr/vista/actions

Spatio-temporal bag-of-features

We use global spatio-temporal grids

- In the spatial domain:
  - 1x1 (standard BoF)
  - 2x2, o2x2 (50% overlap)
  - h3x1 (horizontal), v1x3 (vertical)
  - 3x3
- In the temporal domain:
  - t1 (standard BoF), t2, t3

Quantization:

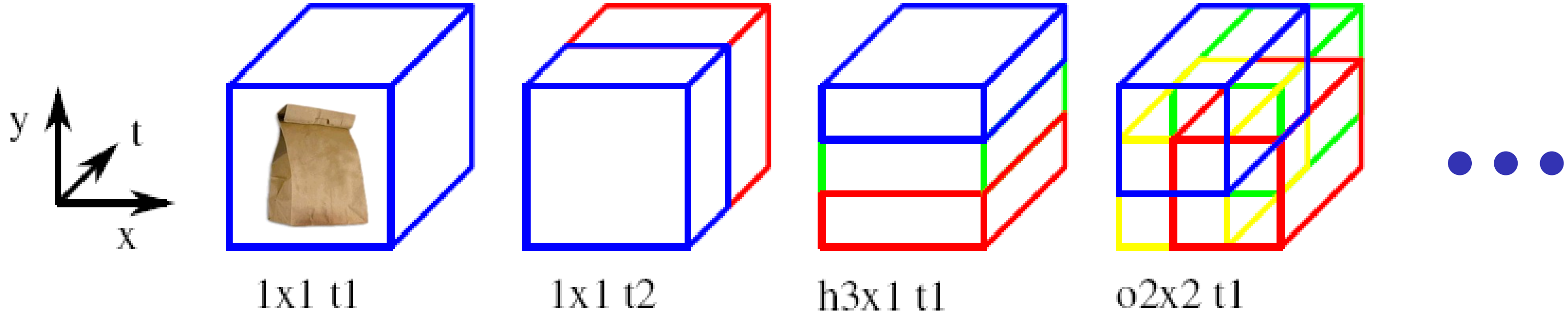


Figure: Examples of a few spatio-temporal grids

Multi-channel chi-square kernel

We use SVMs with a multi-channel chi-square kernel for classification

$$K(H_i, H_j) = \exp\left(-\sum_{c \in \mathcal{C}} \frac{1}{A_c} D_c(H_i, H_j)\right)$$

- Channel c is a combination of a detector, descriptor and a grid
- $D_c(H_i, H_j)$ is the chi-square distance between histograms
- A_c is the mean value of the distances between all training samples
- The best set of channels C for a given training set is found based on a greedy approach

Combining channels

Task	HoG BoF	HoF BoF	Best chan.	Best comb.
KTH multi-class	81.6%	89.7%	91.1%	91.8%
Action AnswerPhone	13.4%	24.6%	26.7%	32.1%
Action GetOutCar	21.9%	14.9%	22.5%	41.5%
Action HandShake	18.6%	12.1%	23.7%	32.3%
Action HugPerson	29.1%	17.4%	34.9%	40.6%
Action Kiss	52.0%	36.5%	52.0%	53.3%
Action SitDown	29.1%	20.7%	37.8%	38.6%
Action SitUp	6.5%	5.7%	15.2%	18.2%
Action StandUp	45.4%	40.0%	45.4%	50.5%

Table: Classification performance of different channels and their combinations

- It is worth trying different grids
- It is beneficial to combine channels

Evaluation of spatio-temporal grids

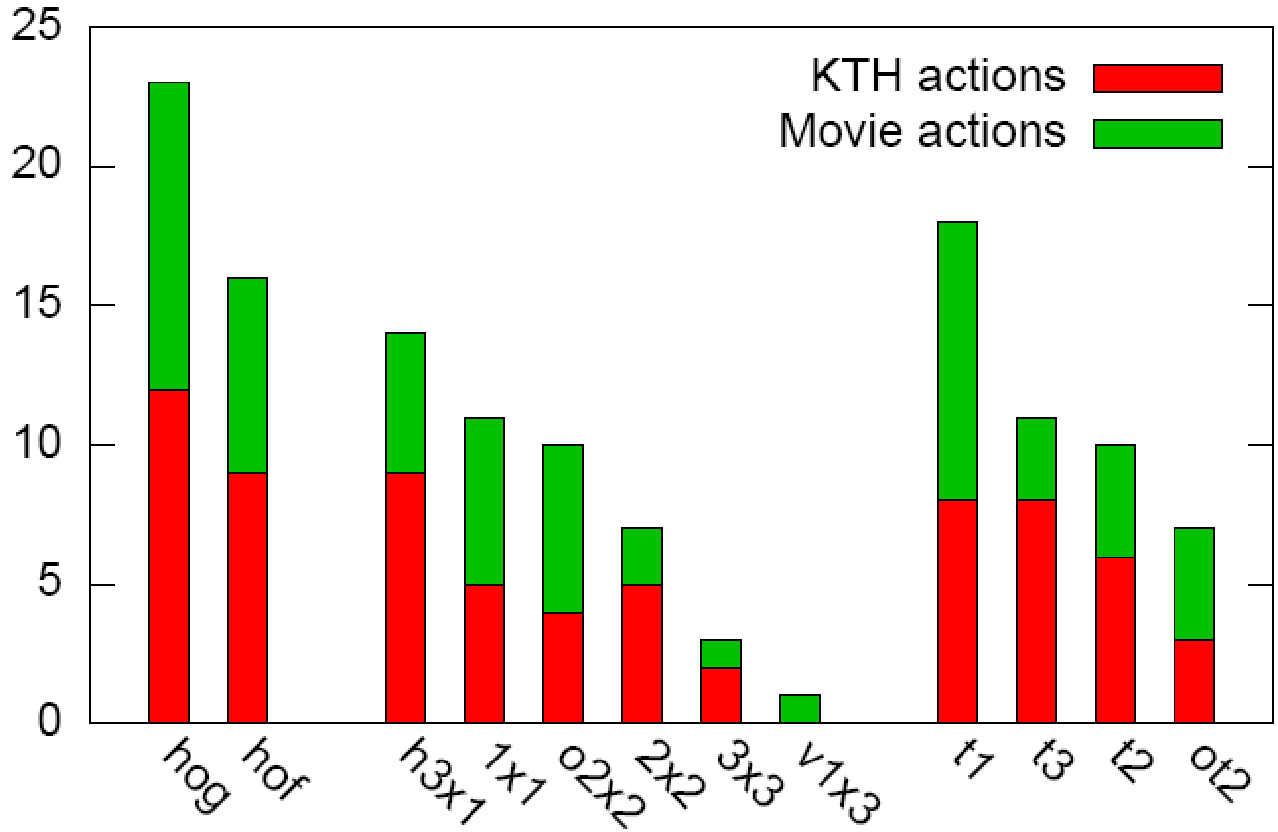


Figure: Number of occurrences for each channel component within the optimized channel combinations for the KTH action dataset and our manually labeled movie dataset

Comparison to the state-of-the-art

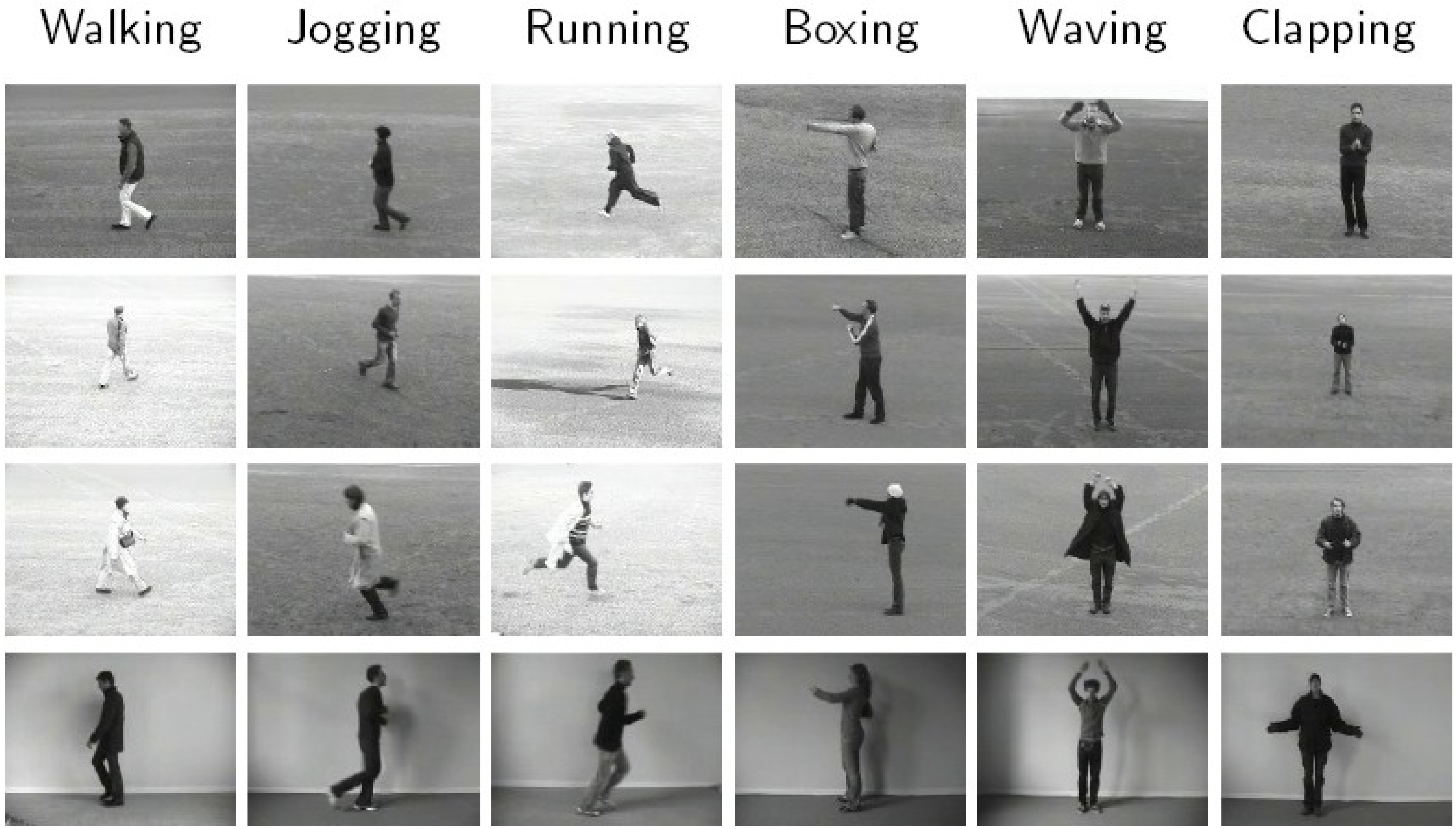


Figure: Sample frames from the KTH actions sequences, all six classes (columns) and scenarios (rows) are presented

Comparison to the state-of-the-art

Method	Schuldt et al.	Niebles et al.	Wong et al.	Nowozin et al.	ours
Accuracy	71.7%	81.5%	86.7%	87.0%	**91.8%**

Table: Average class accuracy on the KTH actions dataset

	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	.99	.01	.00	.00	.00	.00
Jogging	.04	.89	.07	.00	.00	.00
Running	.01	.19	.80	.00	.00	.00
Boxing	.00	.00	.00	.97	.00	.03
Waving	.00	.00	.00	.00	.91	.09
Clapping	.00	.00	.00	.05	.00	.95

Table: Confusion matrix for the KTH actions

Training noise robustness

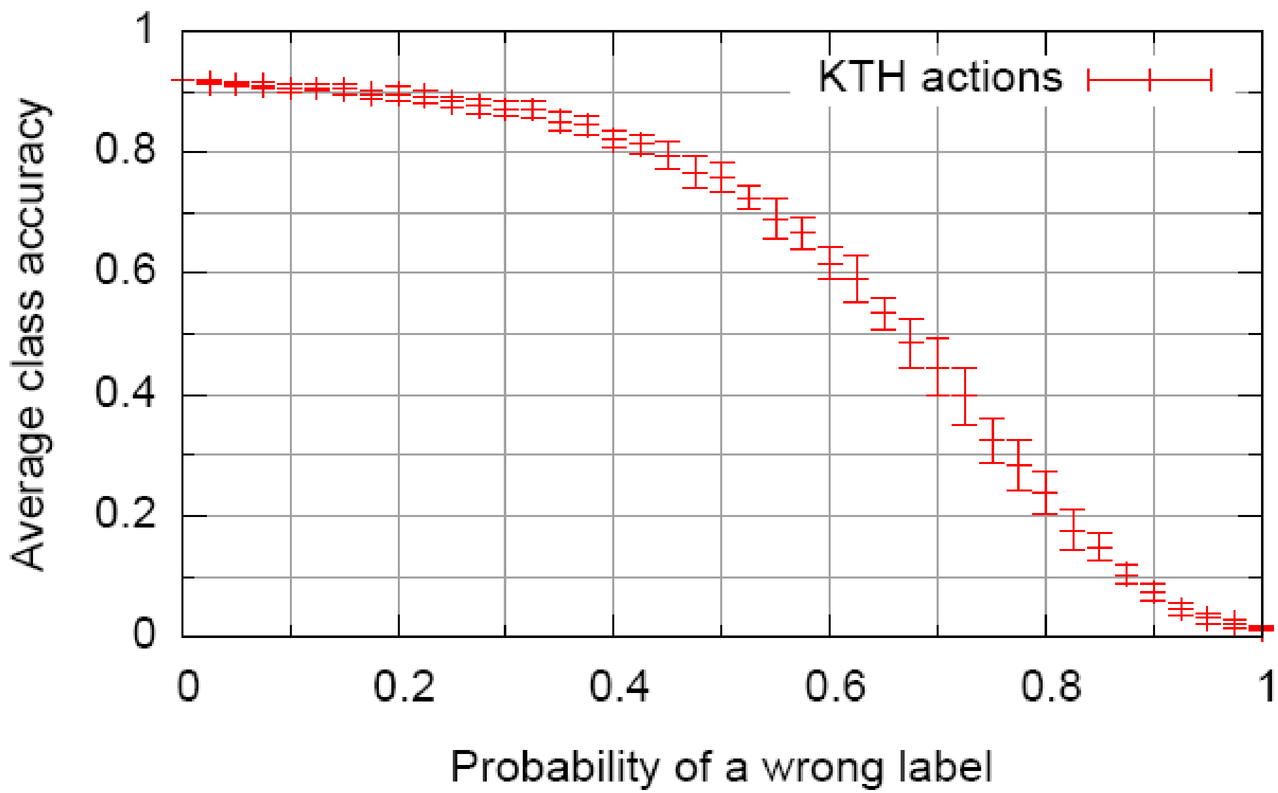


Figure: Performance of our video classification approach in the presence of wrong labels

- Up to *p=0.2* the performance decreases insignificantly
- At *p=0.4* the performance decreases by around 10%

Action recognition in real-world videos

Figure: Example results for action classification trained on the automatically annotated data. We show the key frames for test movies with the highest confidence values for true/false pos/neg

Action recognition in real-world videos

- Note the suggestive FP: hugging or answering the phone
- Note the difficult FN: getting out of car or handshaking

Action recognition in real-world videos

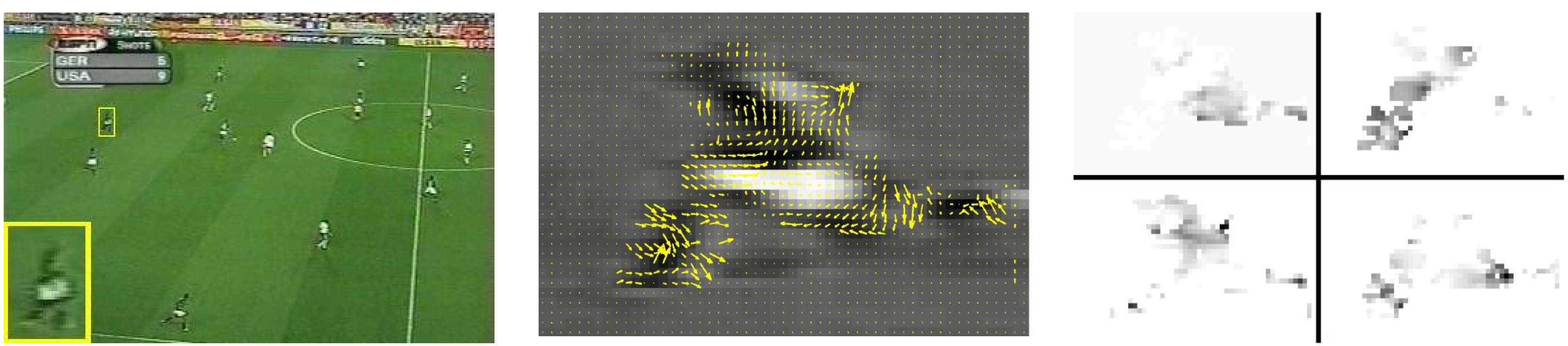
	Clean	Automatic	Chance
AnswerPhone	32.1%	16.4%	10.6%
GetOutCar	41.5%	16.4%	6.0%
HandShake	32.3%	9.9%	8.8%
HugPerson	40.6%	26.8%	10.1%
Kiss	53.3%	45.1%	23.5%
SitDown	38.6%	24.8%	13.8%
SitUp	18.2%	10.4%	4.6%
StandUp	50.5%	33.6%	22.6%

Table: Average precision (AP) for each action class of our test set. We compare results for clean (annotated) and automatic training data. We also show results for a random classifier (chance)

Action recognition in real-world videos

	Clean	Automatic	HoF BoF	Efros et al.	Chance
AnswerPhone	32.1%	16.4%	24.6%	15.0%	10.6%
GetOutCar	41.5%	16.4%	14.9%	0.0%	6.0%
HandShake	32.3%	9.9%	12.1%	26.3%	8.8%
HugPerson	40.6%	26.8%	17.4%	5.9%	10.1%
Kiss	53.3%	45.1%	36.5%	47.6%	23.5%
SitDown	38.6%	24.8%	20.7%	27.3%	13.8%
SitUp	18.2%	10.4%	5.7%	10.0%	4.6%
StandUp	50.5%	33.6%	40.0%	16.7%	22.6%
Average	38.4%	22.9%	21.5%	18.6%	12.5%

Recognizing Action at A Distance A.A. Efros, A.C. Berg, G. Mori and J. Malik



[ICCV 2003]