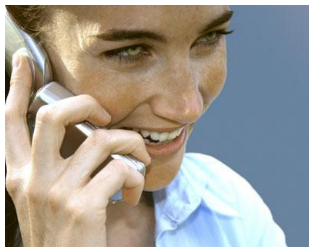
Activity Recognition

Yimeng Zhang

11/29/11

Adapted Slides from Efros' and Laptev's talks

Motivation


- Activity
  - Verb/predicate (object: noun) → human actions
  - Usually detected from a video
- Applications
  - Content-based browsing

    e.g. fast-forward to the next goal scoring scene

    e.g. find "Bush shaking hands with Putin"
  - Video Surveillance

    Monitor the crime related activities
  - Human scientists

    influence of smoking in movies on adolescent smoking

KTH Activity Dataset

UCF Sports Dataset

Hollywood Movie Dataset –v2

Early work: holistic model

- [Efros et al. ICCV 03]
- Tracking the person

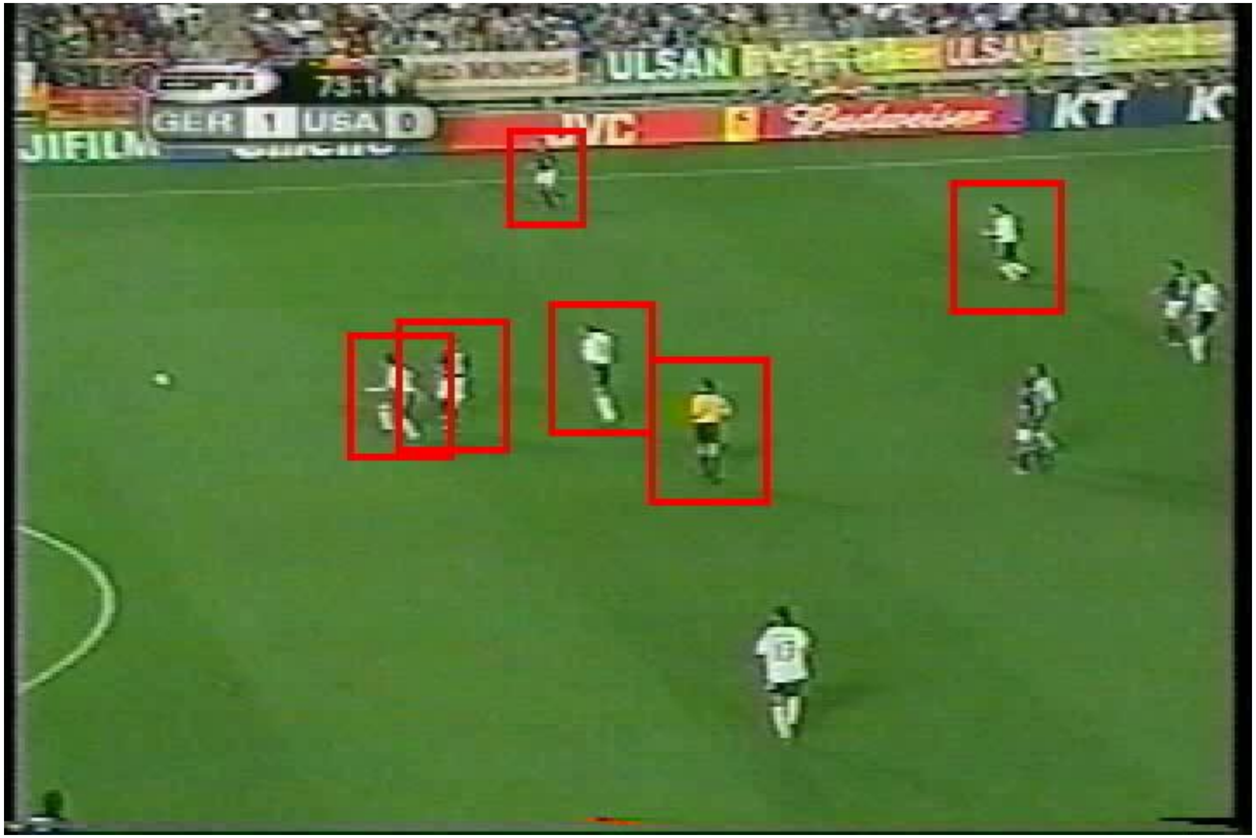
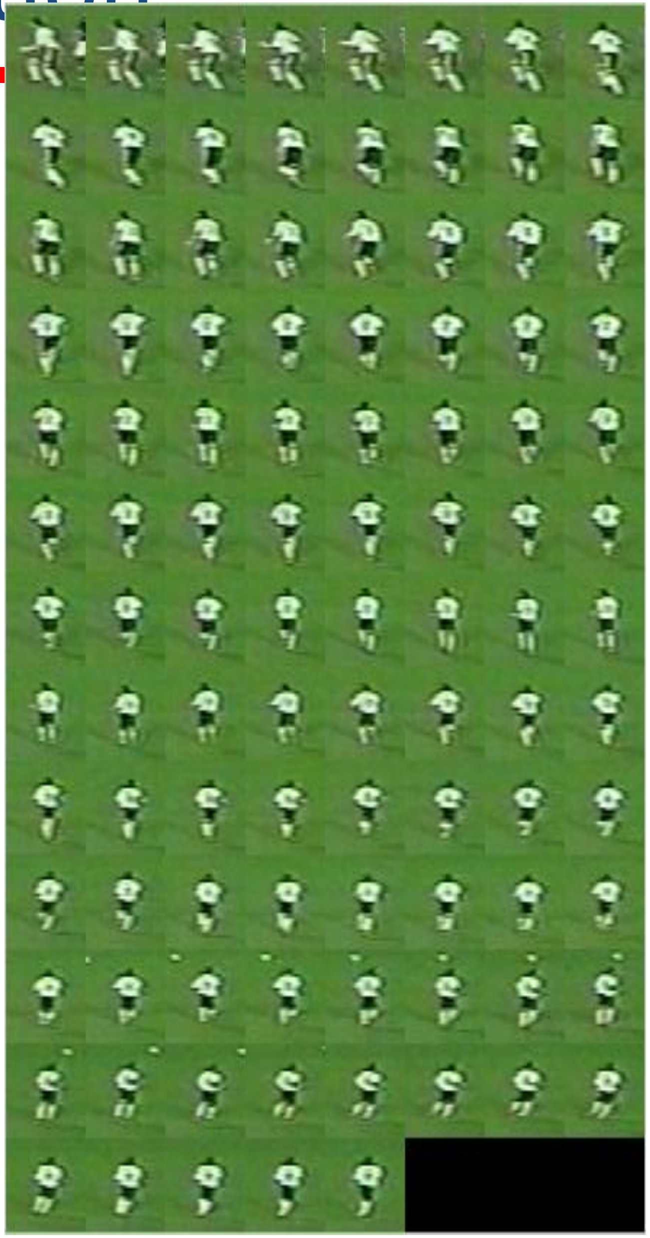


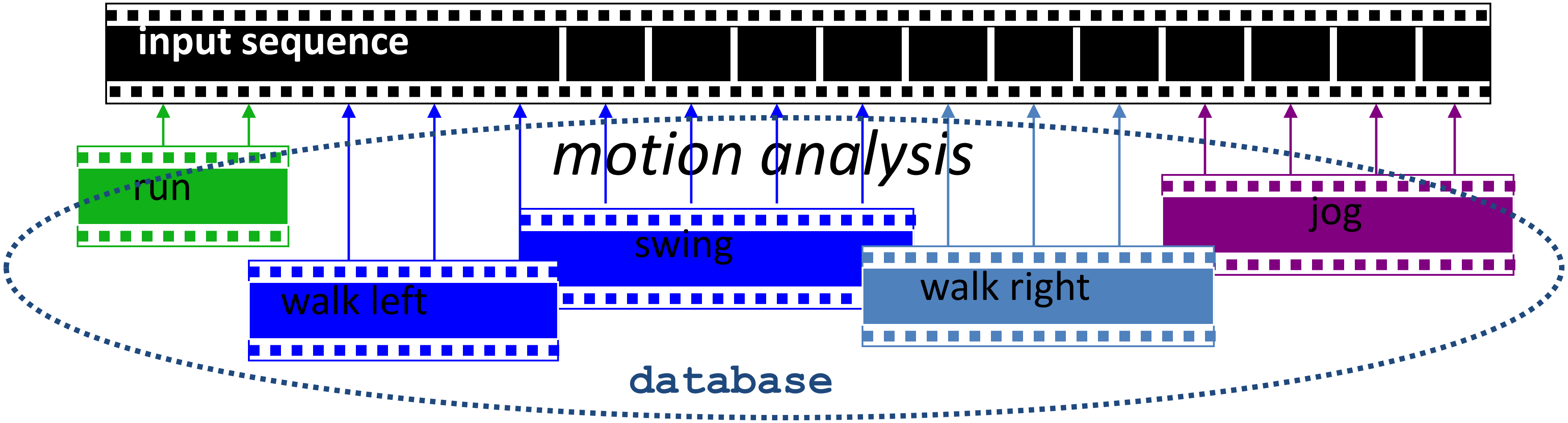
Figure-centric Representation

- Stabilized spatio-temporal volume
  - No translation information
  - All motion caused by person's limbs



Remembrance of Things Past

- "Explain" novel motion sequence by matching to previously seen video clips
  - For each frame, match based on some temporal extent



<u>Challenge:</u> how to compare motions?

Spatial Motion Descriptor

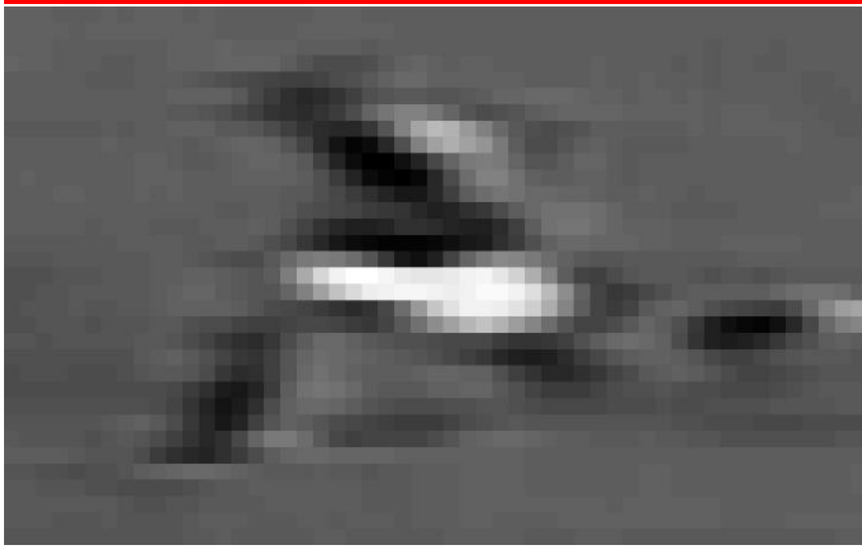
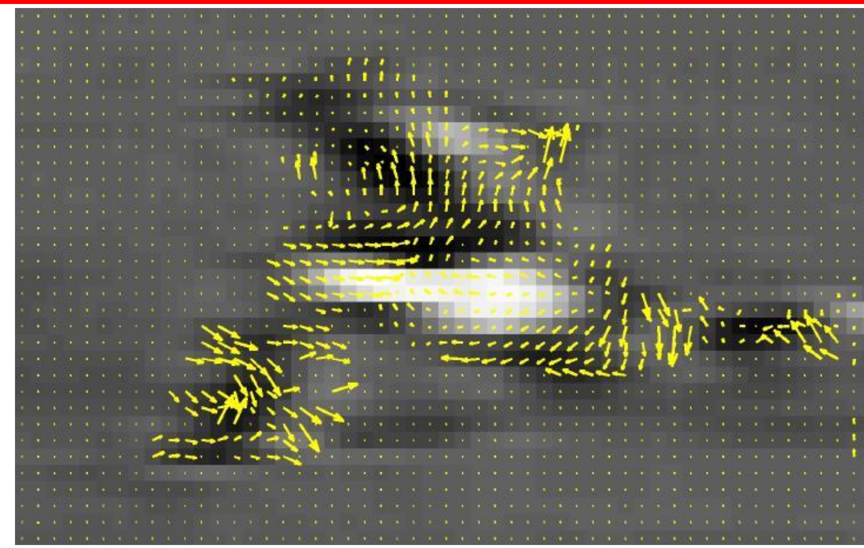
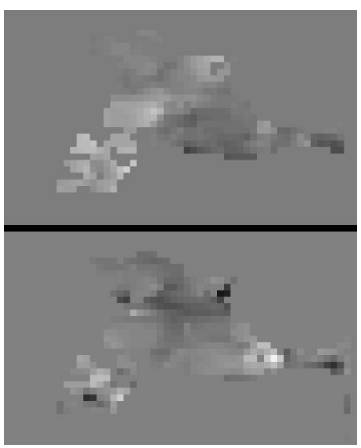


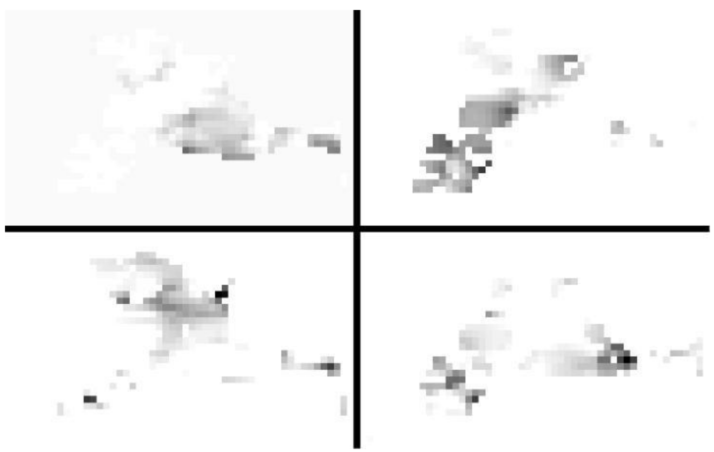
Image frame



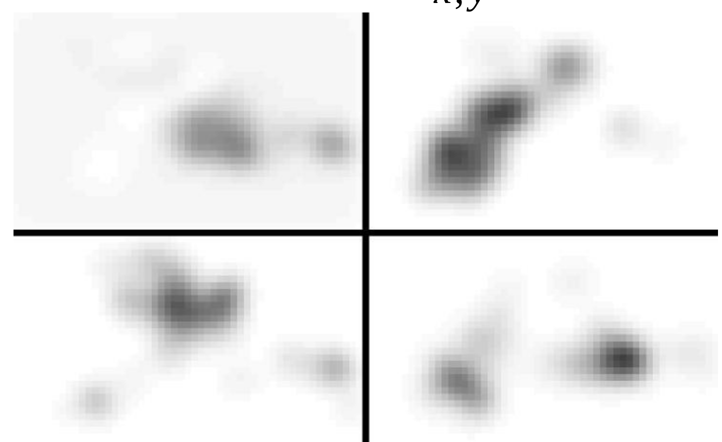
Optical flow $F_{x,y}$



F_x, F_y



$F_x^-, F_x^+, F_y^-, F_y^+$



blurred $F_x^-, F_x^+, F_y^-, F_y^+$

Football Actions: matching

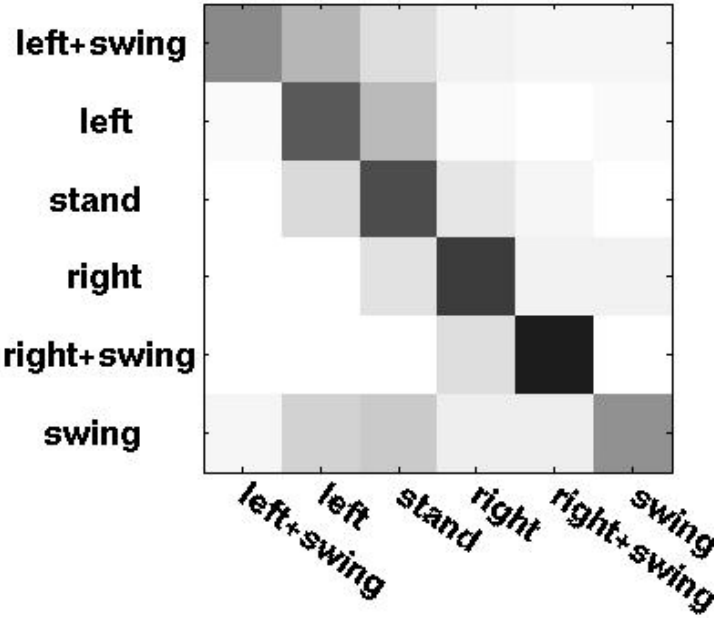
Input Sequence

Matched Frames

input matched

Classifying Tennis Actions

6 actions; 4600 frames; 7-frame motion descriptor
Woman player used as training, man as testing.



Holistic Model

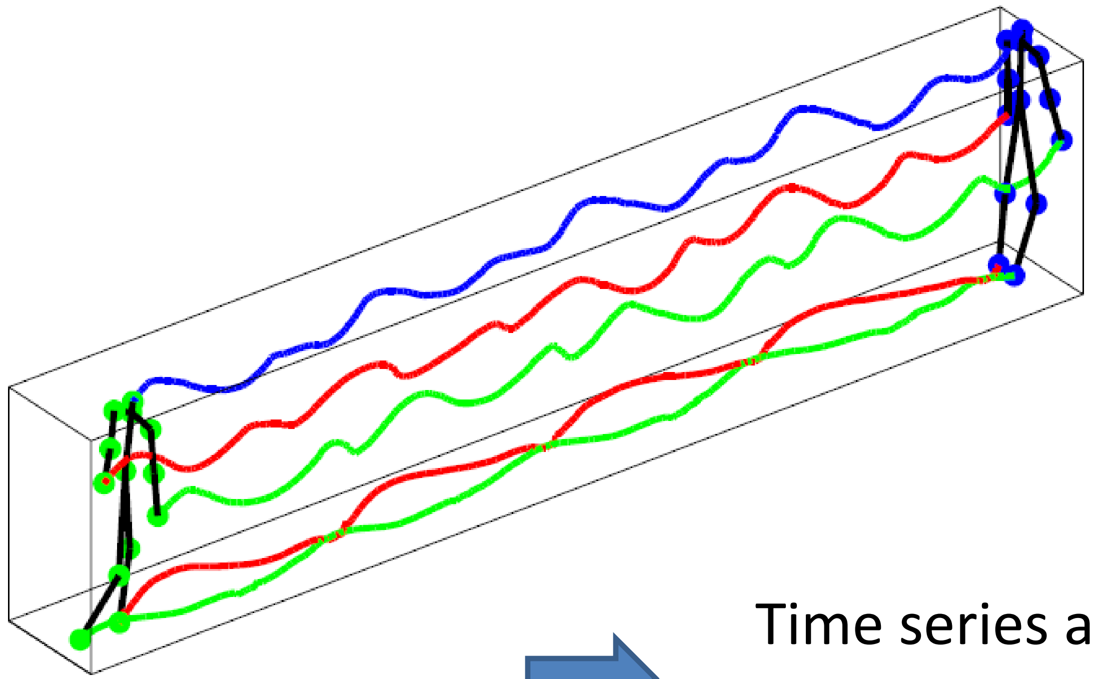
- Advantage
  - Rich spatial modeling of different body parts
  - High discrimination

- Disadvantage
  - Require background subtraction
  - Require robust person tracking
  - Does not generalize well

Body Part based Model

- [Ali et al. ICCV 07]

Body Part Trajectories



Time series analysis of the trajectories

Problems of Holistic and Body Part Methods

Holistic or Body Part Methods:

- Camera stabilization
- Segmentation
- Tracking

Common problems:

- Complex & changing BG
- Appearance of new OBJ

[Laptev et al. CVPR08]

Activity Dataset “in the wild”

Laptev et al. CVPR 2008

What are human actions?

- Actions in current datasets:

KTH action dataset

- Actions “In the Wild”:

Actions in movies

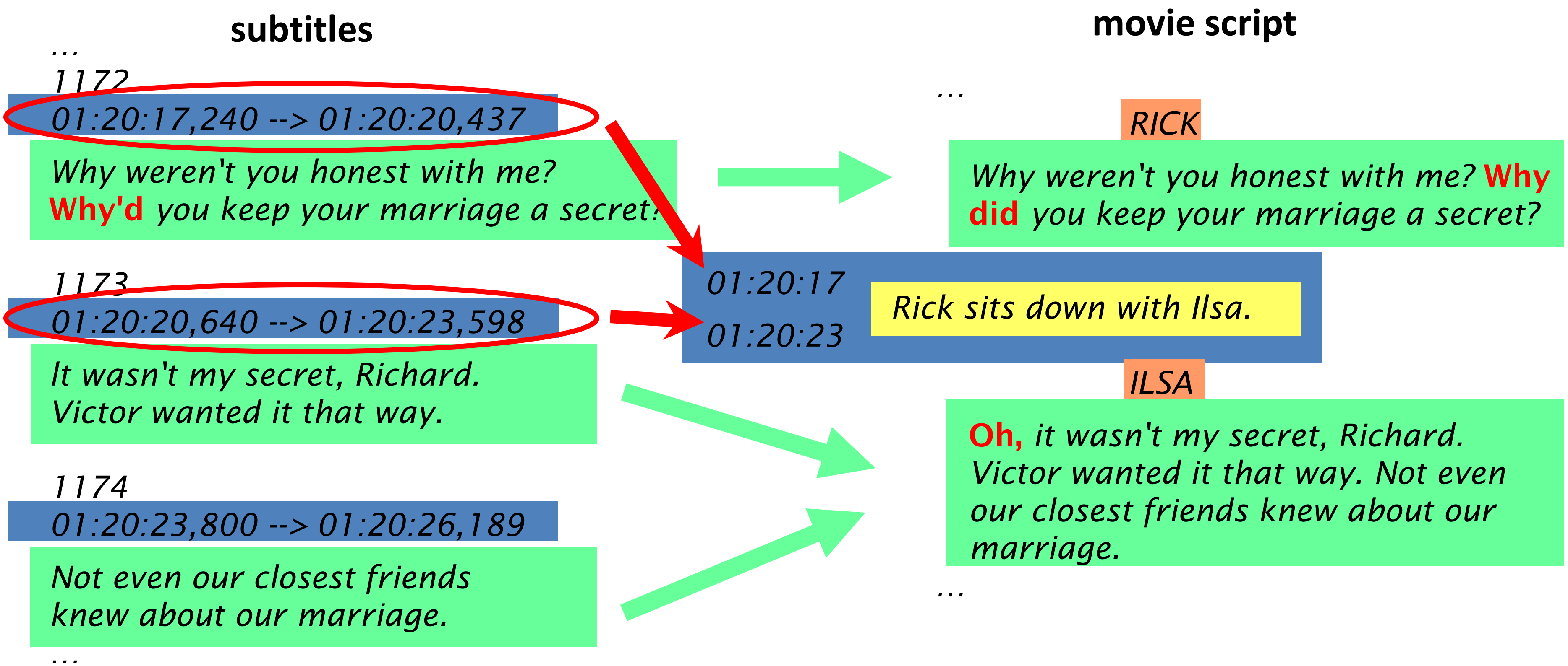
- Realistic variation of human actions
- Many classes and many examples per class

Problems:

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

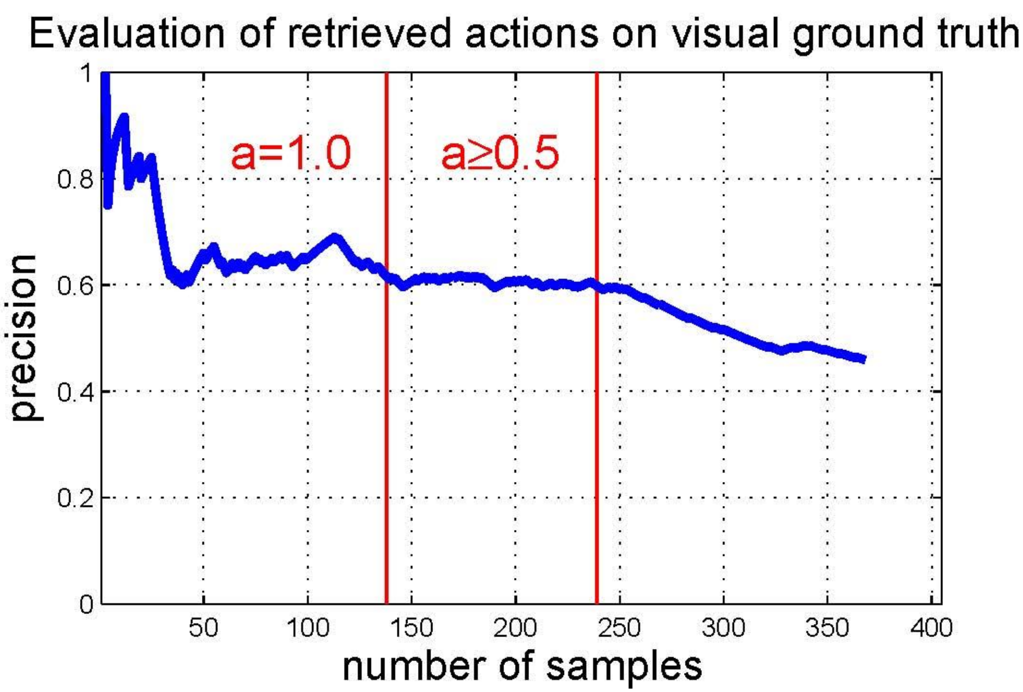
Automatic video annotation using scripts

- 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 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”

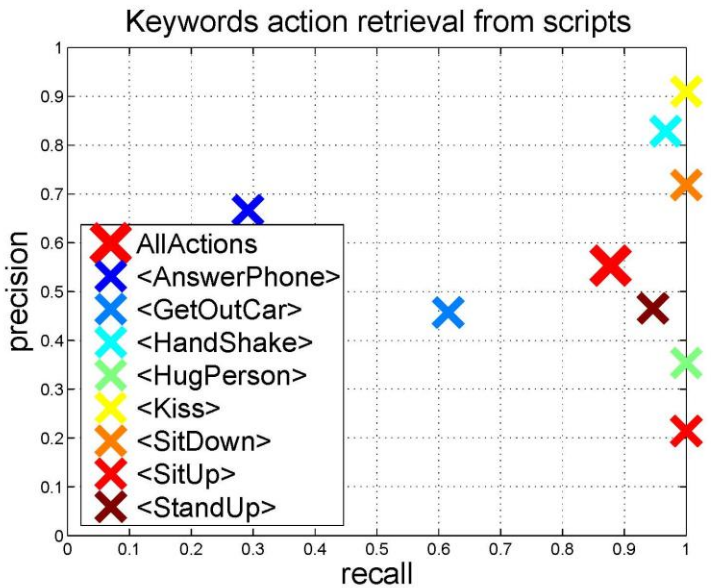
A black car pulls up, two army officers get out.

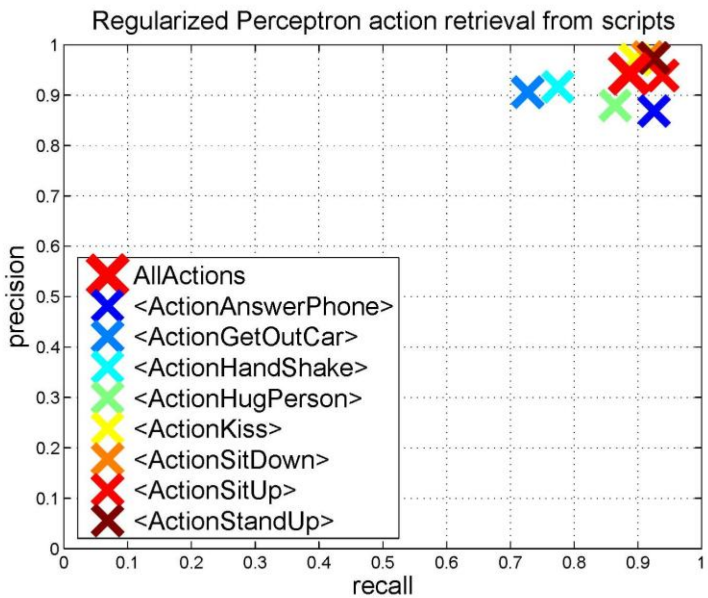
Text-based action retrieval

- 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

	<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

12 movies

23	13	19	22	51	30	10	49	217

test set

20 different movies

Hollywood-2 dataset: >1700 video clips of 12 categories

Training noise robustness

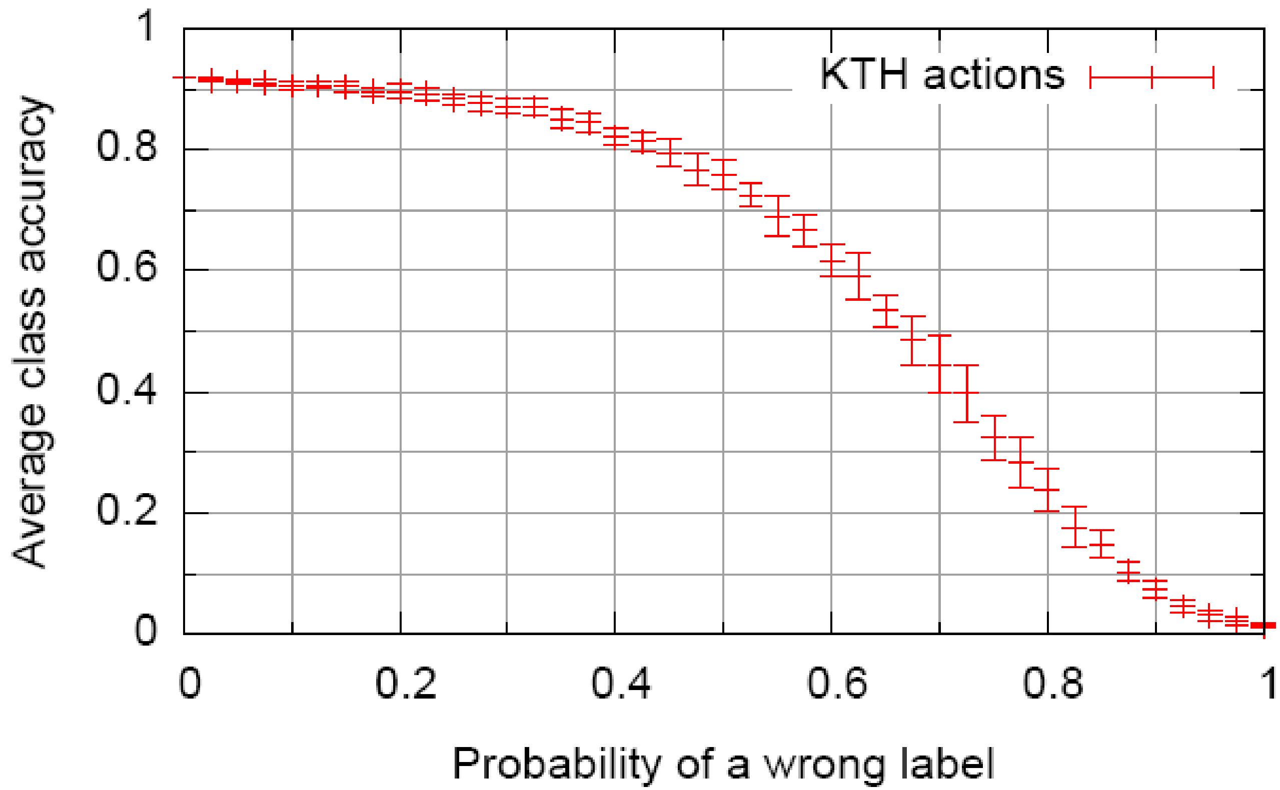


Figure: Performance of our video classification approach in the presence of wrong labels

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

Hollywood Movie Result

	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)

[Laptev et al. CVPR08]

Space-Time Local Features

[Dollar et al. PETS Workshop 2005]

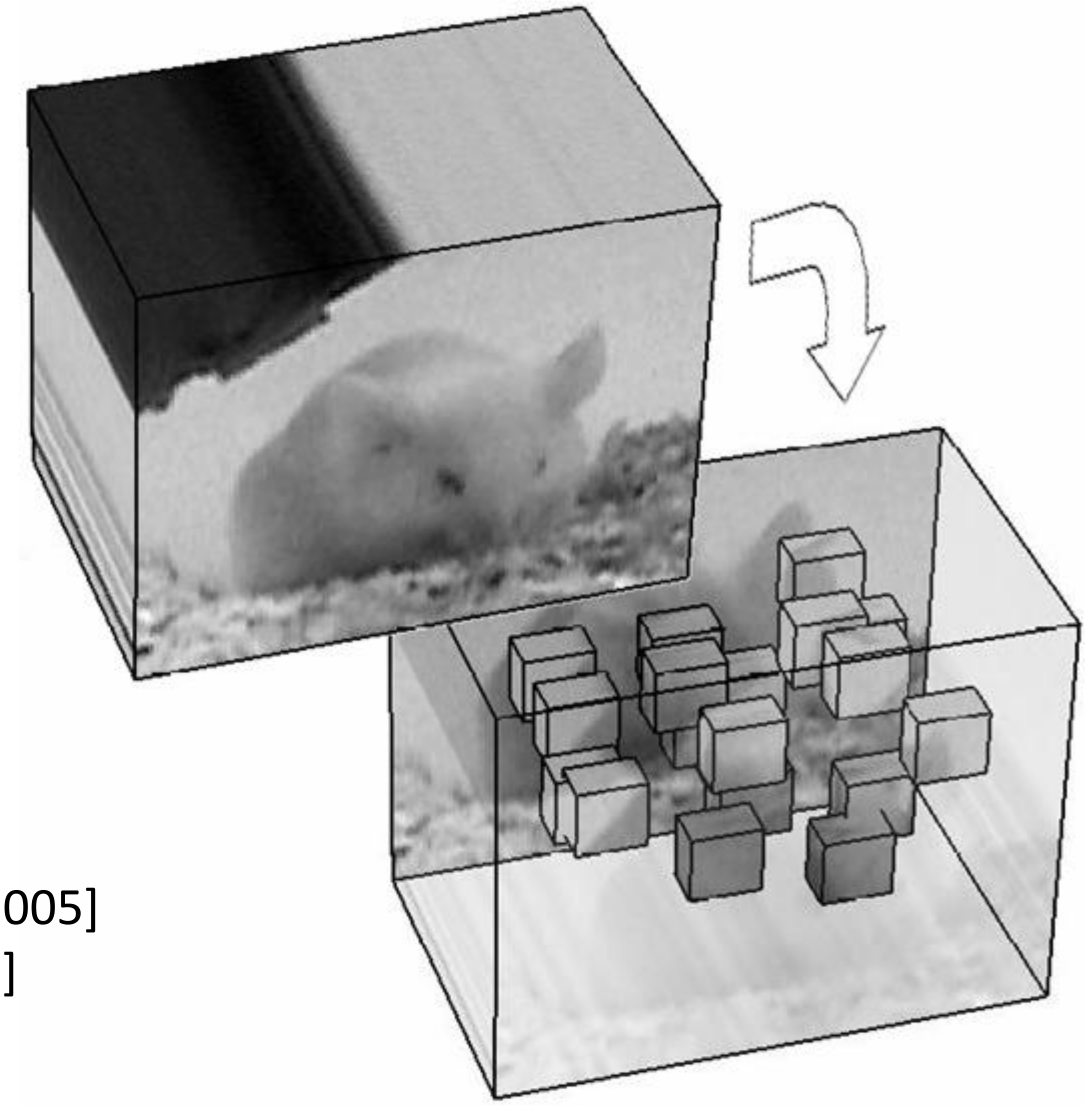
[Laptev et al. ICCV 03, CVPR08]

[Willems et al. ECCV 08]

[Wang et al. BMVC 09]

Space-time Local Features

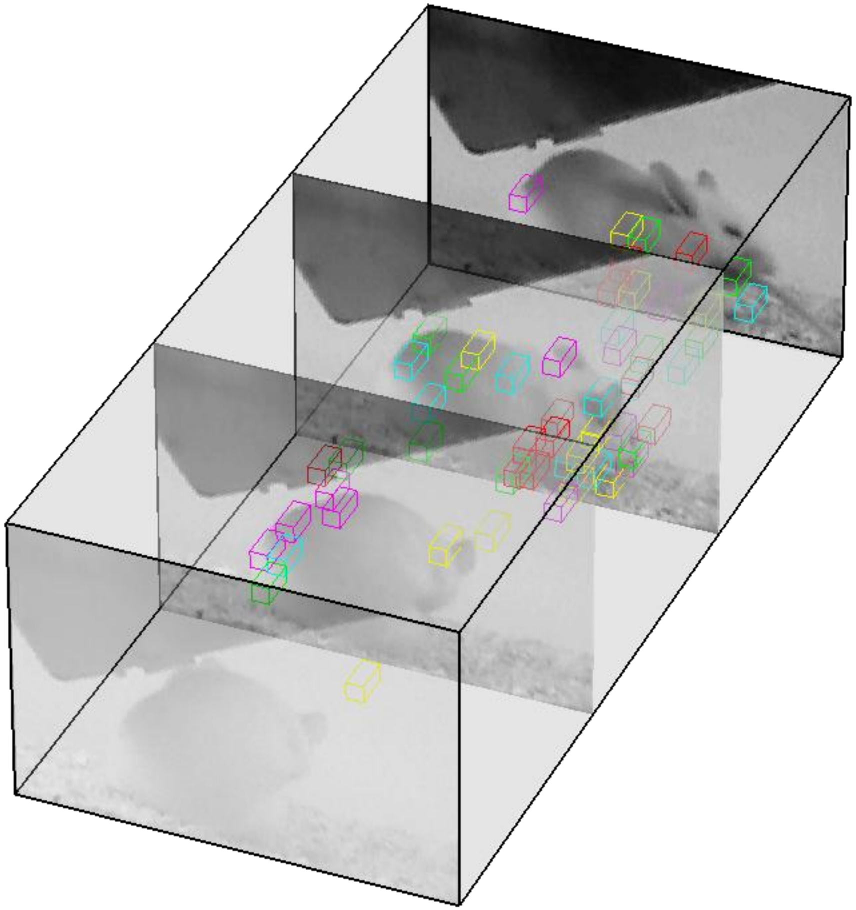
Consider local spatio-temporal neighborhoods



[Dollar et al. PETS Workshop 2005]
[Laptev et al. ICCV 03, CVPR08]
[Wang et al. BMVC 09]

2D→3D Local Features

- Motivation:
  - *Sparse feature points* extended to the spatio-temporal case



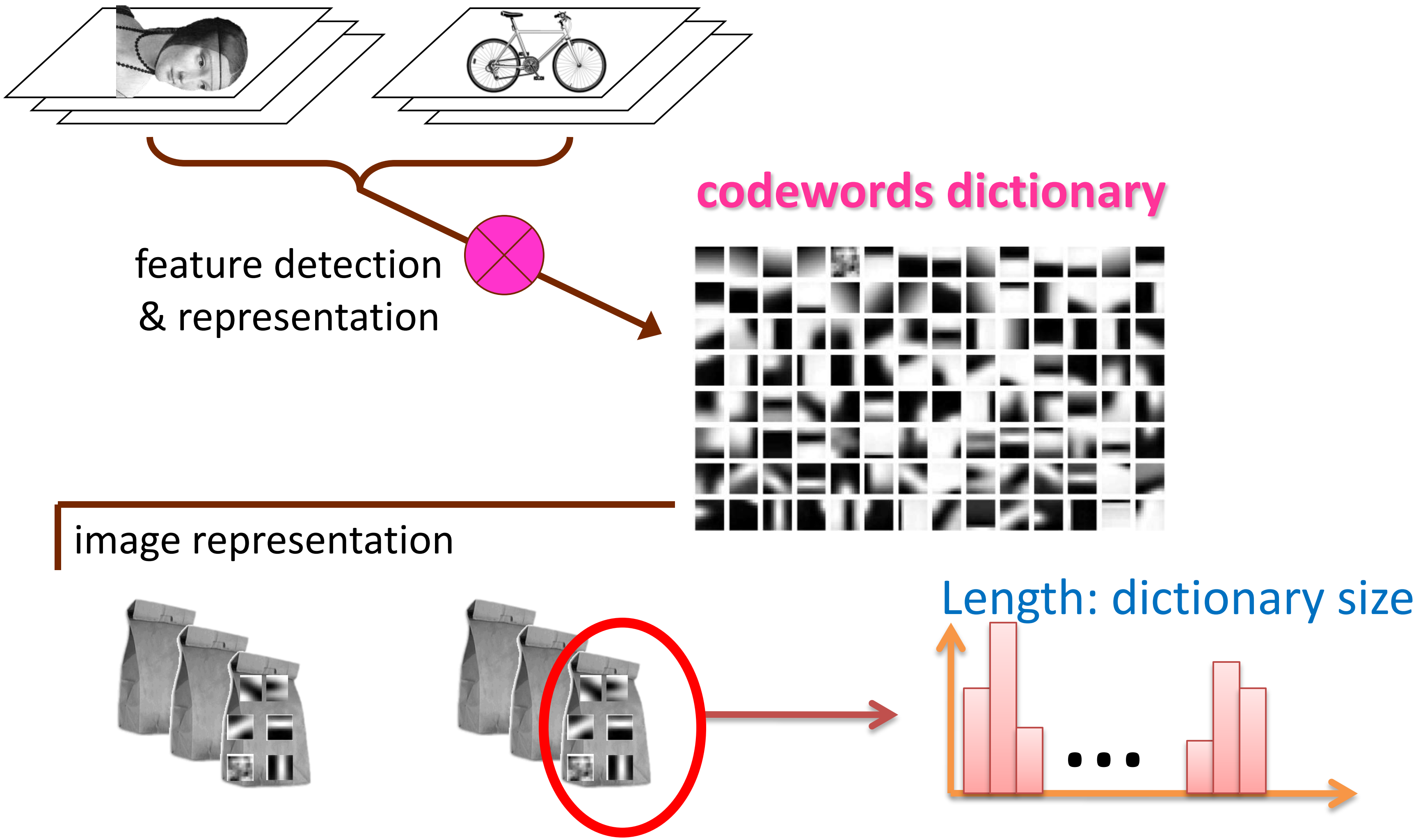
Object Recognition

Advantages of Sparse Features

- Robustness
- Very good results

example from: http://www.robots.ox.ac.uk/~fergus/research/index.html

Bag-of-Words for Object Recognition



Adopted figures from slides of Feifei, Li

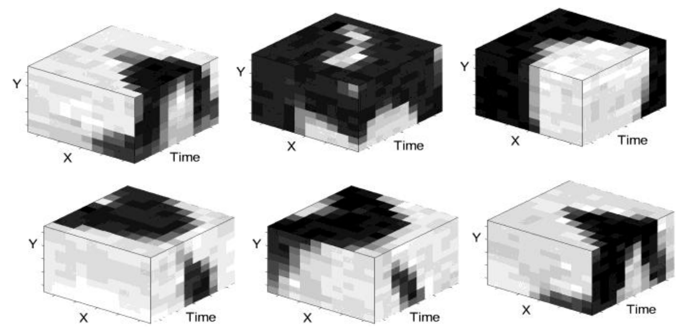
Space-time Bag-of-Words

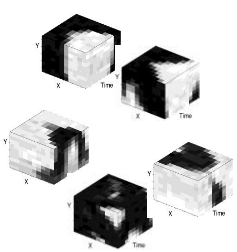
Bag of space-time features + multi-channel SVM

[Schuldt’04, Niebles’06, Zhang’07]

Feature detector

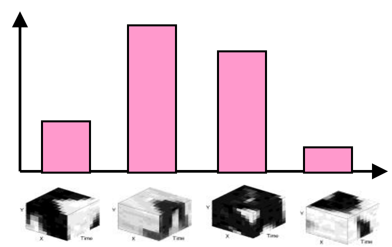
Collection of space-time patches





Descriptors

Histogram of visual words



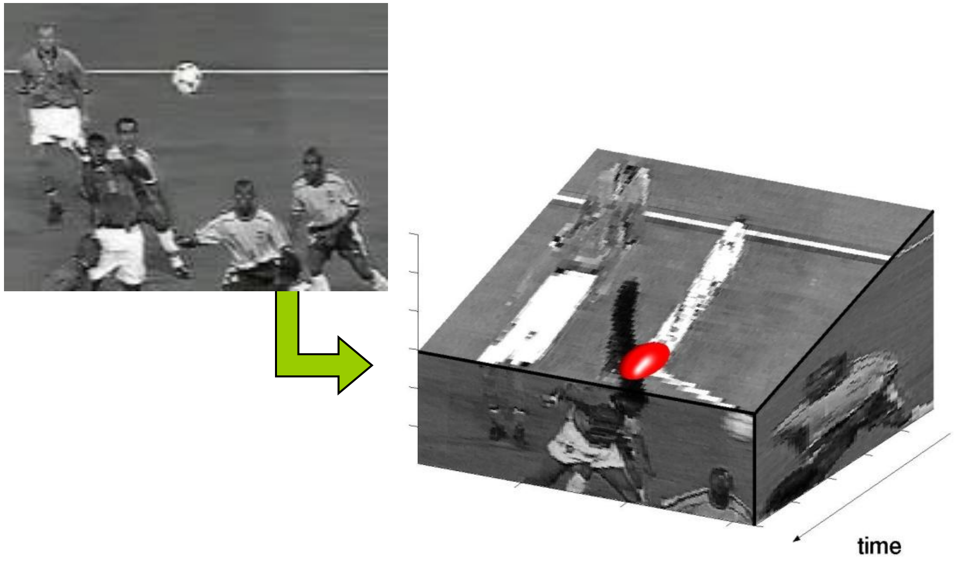
Multi-channel
SVM
Classifier

Space-Time Feature: Detector (1)

- Harris-3D
  - [Laptev et al., ICCV 03]
  - Space-time corner detector

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

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



Space-Time Feature: Detector (2)

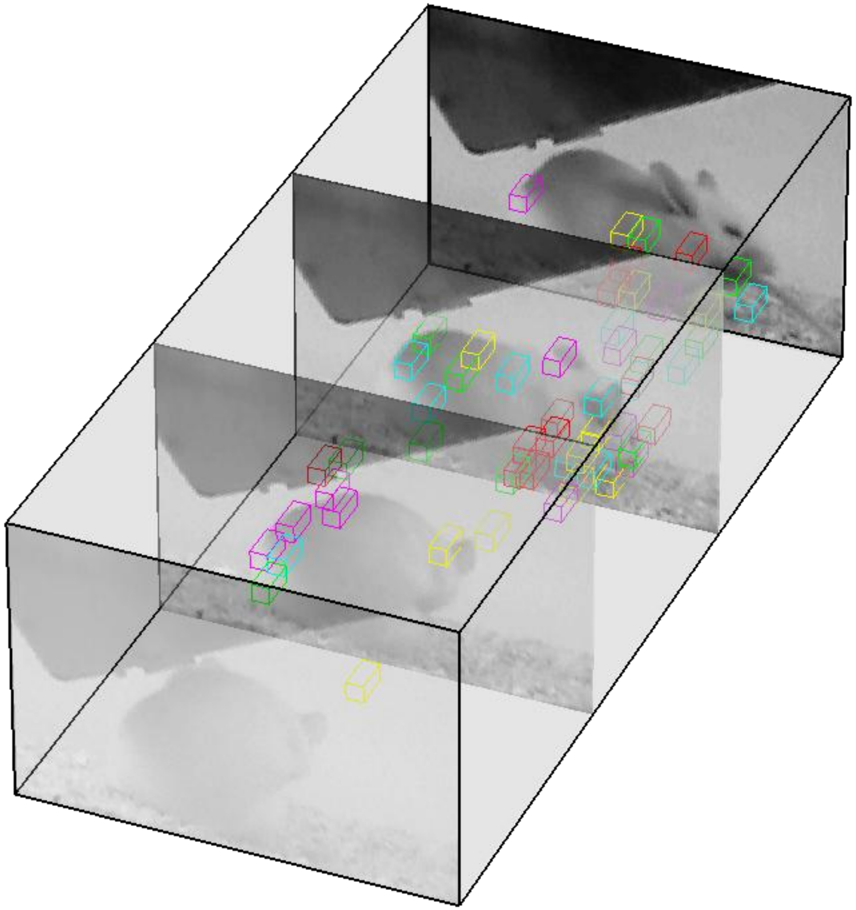
- Cuboids
  - [Dollar et al. 2005]
- Response Function

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2$$

  - Spatial Filter: Gaussian
  - Temporal Filter: Gabor

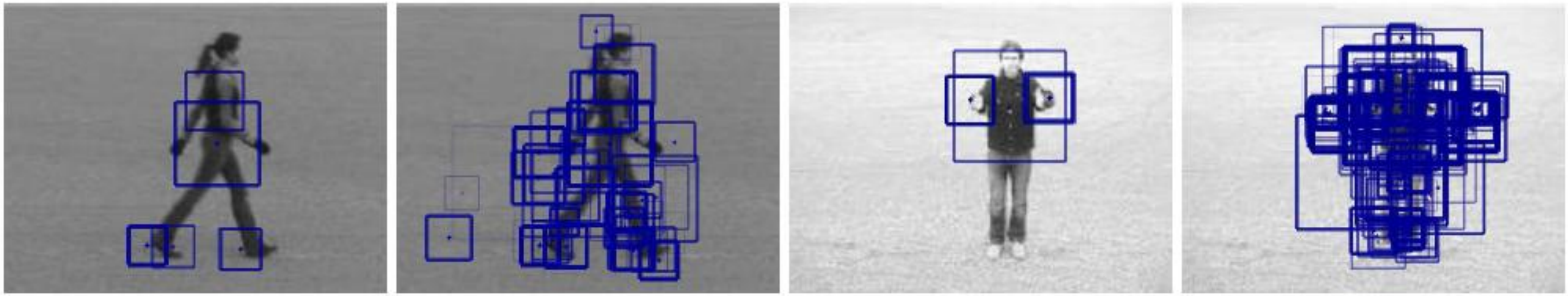
$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t \omega) e^{-t^2/\tau^2}$$

$$h_{od}(t; \tau, \omega) = -\sin(2\pi t \omega) e^{-t^2/\tau^2}$$



Space-Time Feature: Detector (3)

- Hessian – 3D
  - [Willems et al. ECCV 08]
  - Blob-like features



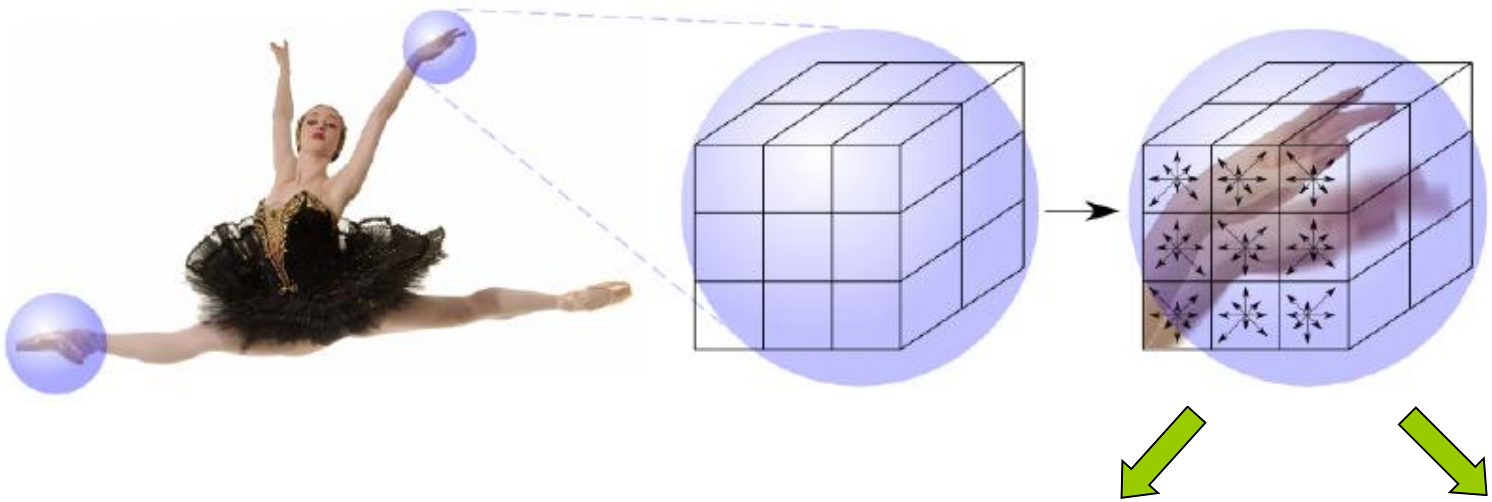
- Dense sampling [Wang et al. BMVC 09]

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

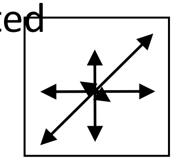
Space-Time Feature: Detector (4)

Space-Time Features: Descriptor

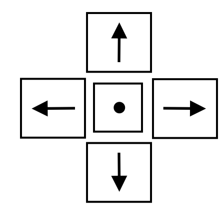
Multi-scale space-time patches from corner detector



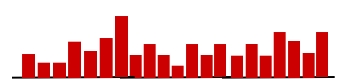
Histogram of oriented spatial grad. (HOG)



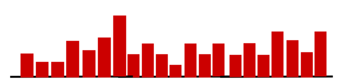
Histogram of optical flow (HOF)



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



3x3x2x4bins **HOG** descriptor



3x3x2x5bins **HOF** descriptor

Comparison - KTH

2391 video clips of 6 categories

walking running jogging handwaving handclapping boxing

	HOG3D	HOG/HOF	HOG	HOF	Cuboids	ESURF
Harris3D	89.0%	**91.8%**	80.9%	**92.1%**	–	–
Cuboids	**90.0%**	88.7%	82.3%	88.2%	89.1%	–
Hessian	84.6%	88.7%	77.7%	88.6%	–	81.4%
Dense	85.3%	86.1%	79.0%	88.0%	–	–

[Wang et al. BMVC 09]

Comparison – UCF Sport

150 video clips of 10 categories


Diving


Kicking


Walking


Skateboarding


High-Bar-Swinging

	HOG3D	HOG/HOF	HOG	HOF	Cuboids	ESURF
Harris3D	79.7%	78.1%	71.4%	75.4%	–	–
Cuboids	**82.9%**	77.7%	72.7%	76.7%	76.6%	–
Hessian	79.0%	79.3%	66.0%	75.3%	–	77.3%
Dense	**85.6%**	81.6%	77.4%	**82.6%**	–	–

[Wang et al. BMVC 09]

Comparison – Hollywood-2 Dataset

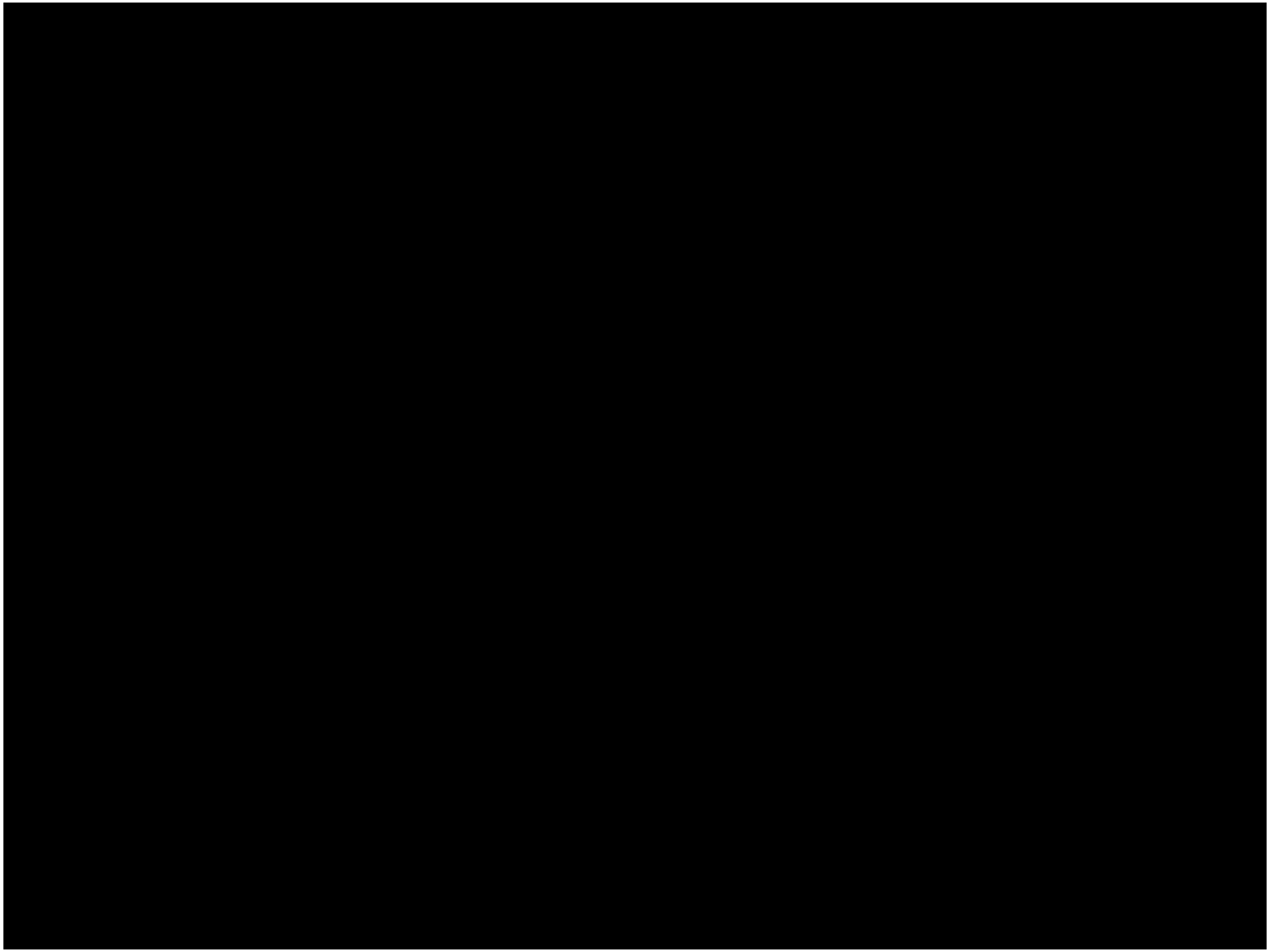
1707 video clips of 12 actions

AnswerPhone GetOutCar HandShake HugPerson Kiss

	HOG3D	HOG/HOF	HOG	HOF	Cuboids	ESURF
Harris3D	43.7%	45.2%	32.8%	43.3%	–	–
Cuboids	45.7%	**46.2%**	39.4%	42.9%	45.0%	–
Hessian	41.3%	**46.0%**	36.2%	43.0%	–	38.2%
Dense	45.3%	**47.4%**	39.4%	45.5%	–	–

[Wang et al. BMVC 09]

Hollywood Movie Demo by [Laptev 08]



Long-range Spatio-Temporal Information

Spatio-temporal Pyramid Matching

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

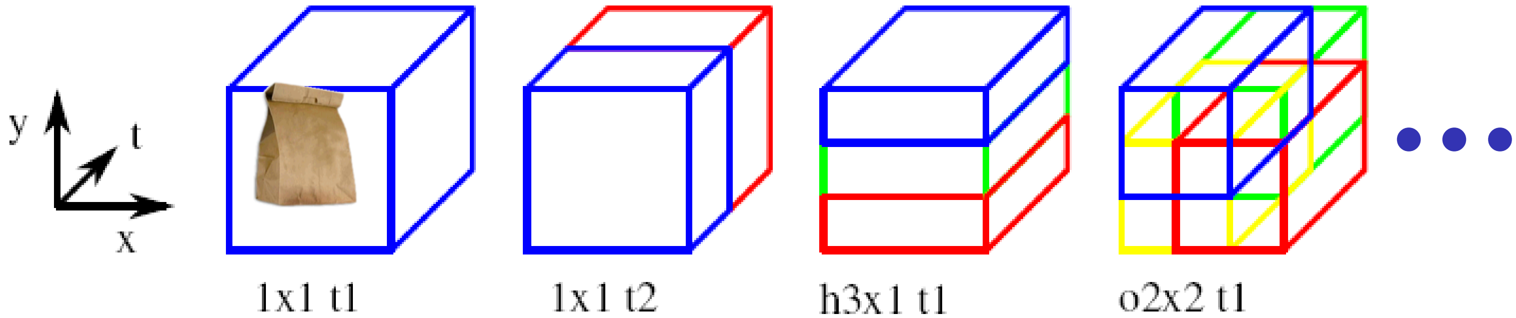


Figure: Examples of a few spatio-temporal grids

[Laptev et al. CVPR08]

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
- 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%

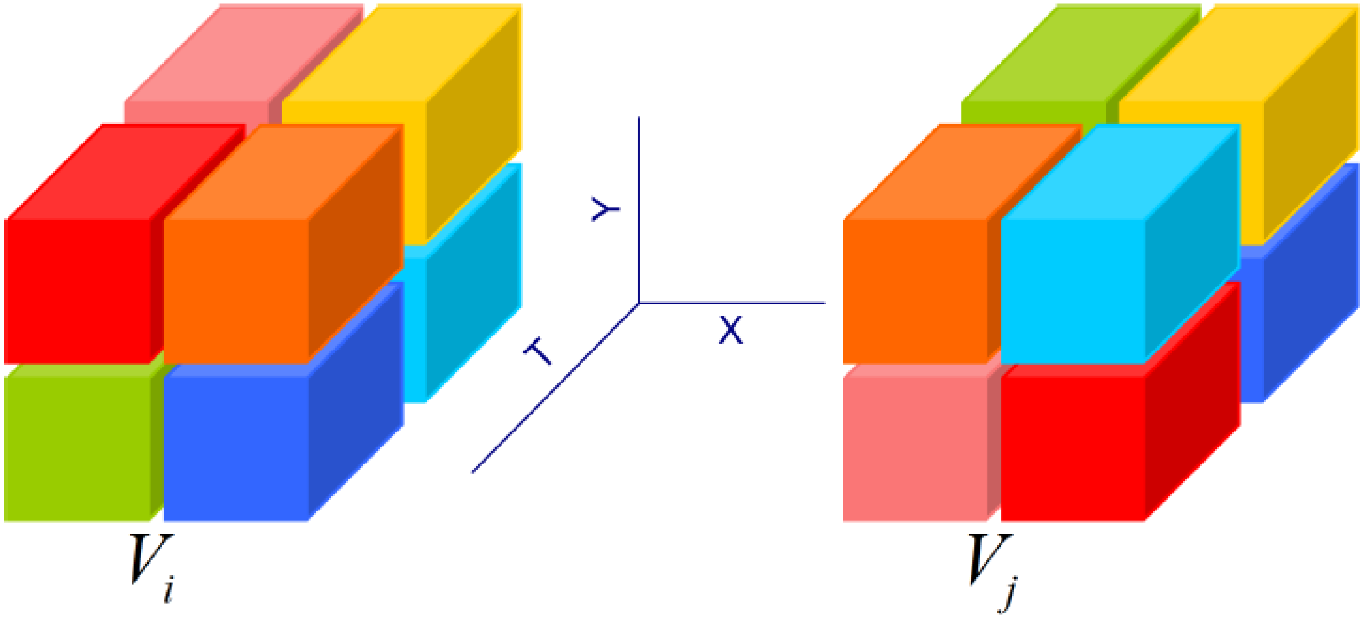
Table: Classification performance of different channels and their combinations

→

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

Aligned Space-Time Pyramid Matching

[Duan et al. CVPR 10]



Find best matching of a binary graph

Aligned Pyramid Matching Result

Table 1. Means and standard deviations (%) of MAPs at different levels using SVM with the default kernel parameter for SIFT features.

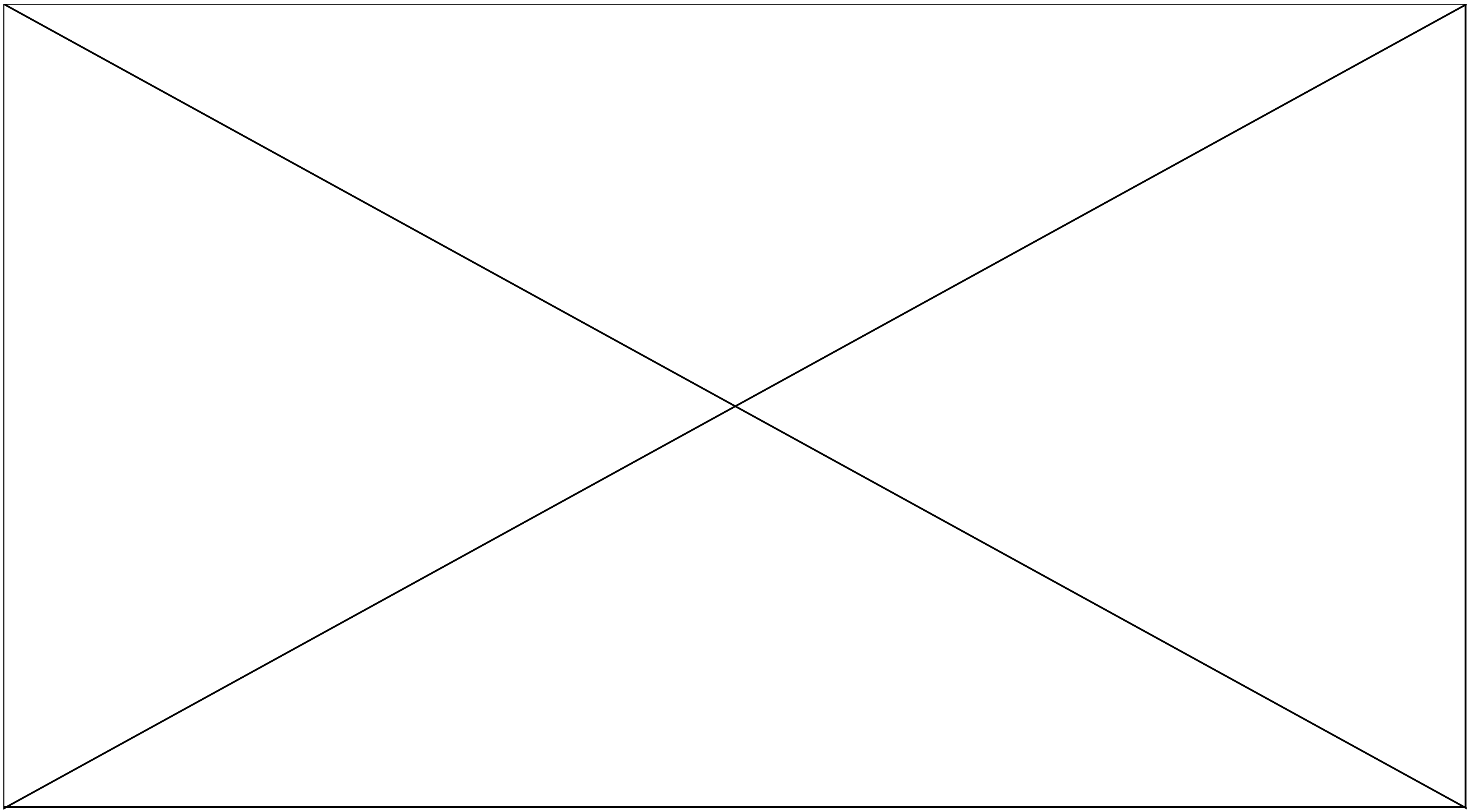
	Gaussian	**Laplacian**	**ISD**	**ID**
Level-0	41.4 ± 3.7	44.2 ± 3.8	45.0 ± 3.5	46.2 ± 4.0
Level-1 (Unaligned)	43.0 ± 2.7	47.7 ± 1.7	49.0 ± 1.6	48.2 ± 1.5
Level-1 (Aligned)	50.4 ± 3.7	53.8 ± 1.8	52.9 ± 3.6	51.0 ± 2.5

Kodak Event Dataset: 1358 video clips of 6 Events

Trajectory of Local Features

- [Messing et al. ICCV 09]
- Track local space-time features

Feature Flow



Cluster Feature Flow

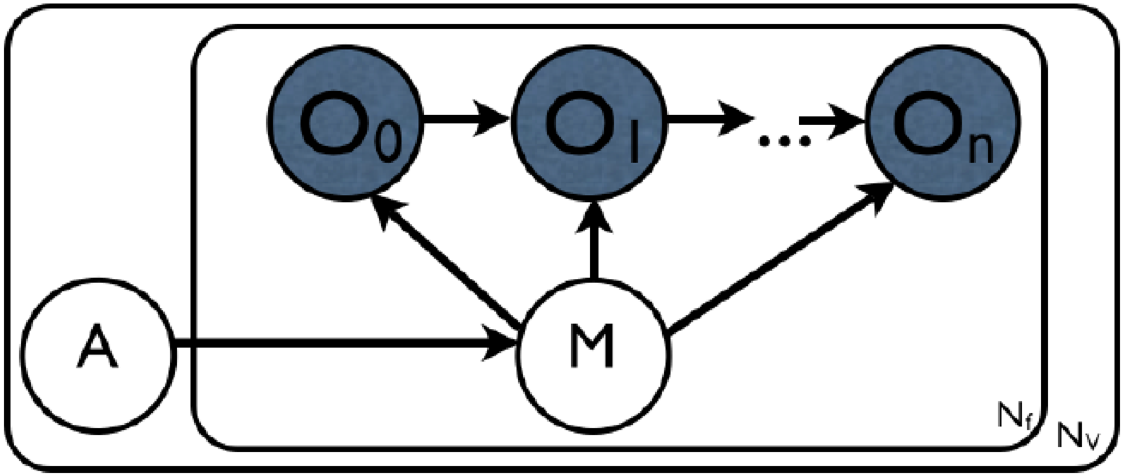
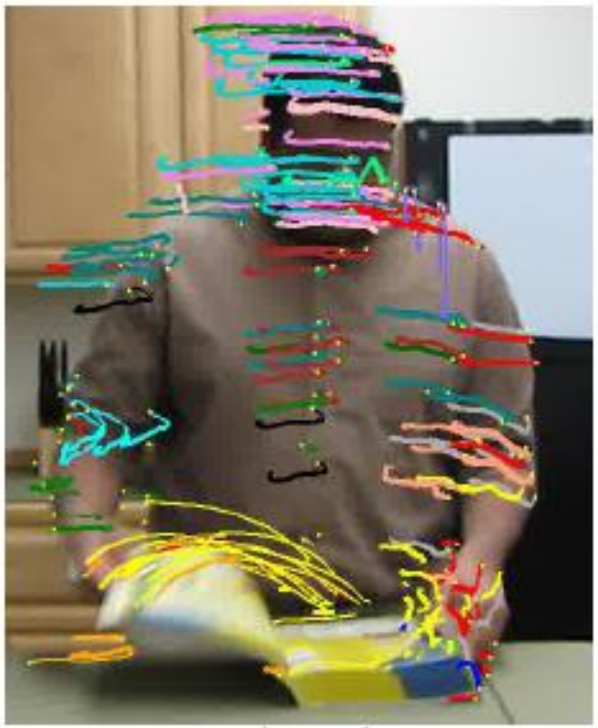
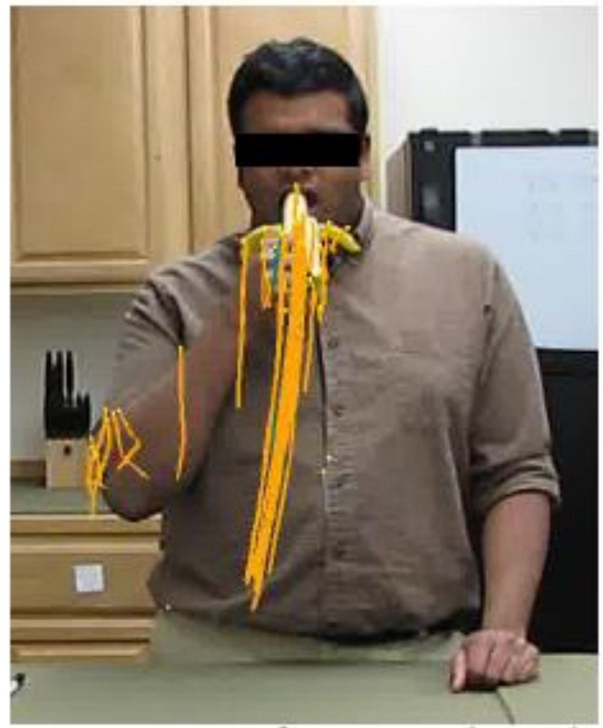


Figure 1. Graphical model for our tracked keypoint velocity history model (Dark circles denote observed variables).

$$
\begin{aligned}
P(A, O) &= \sum_{M} P(A, M, O) = \\
& P(A) \prod_{f}^{N_f} \sum_{i}^{N_m} P(M_f^i | A) P(O_{0,f} | M_f^i) \\
& \prod_{t=1}^{T_f} P(O_{t,f} | O_{t-1,f}, M_f^i) \qquad (5)
\end{aligned}
$$

Trajectory words





Trajectory of Local Features Result

- Daily Living Dataset

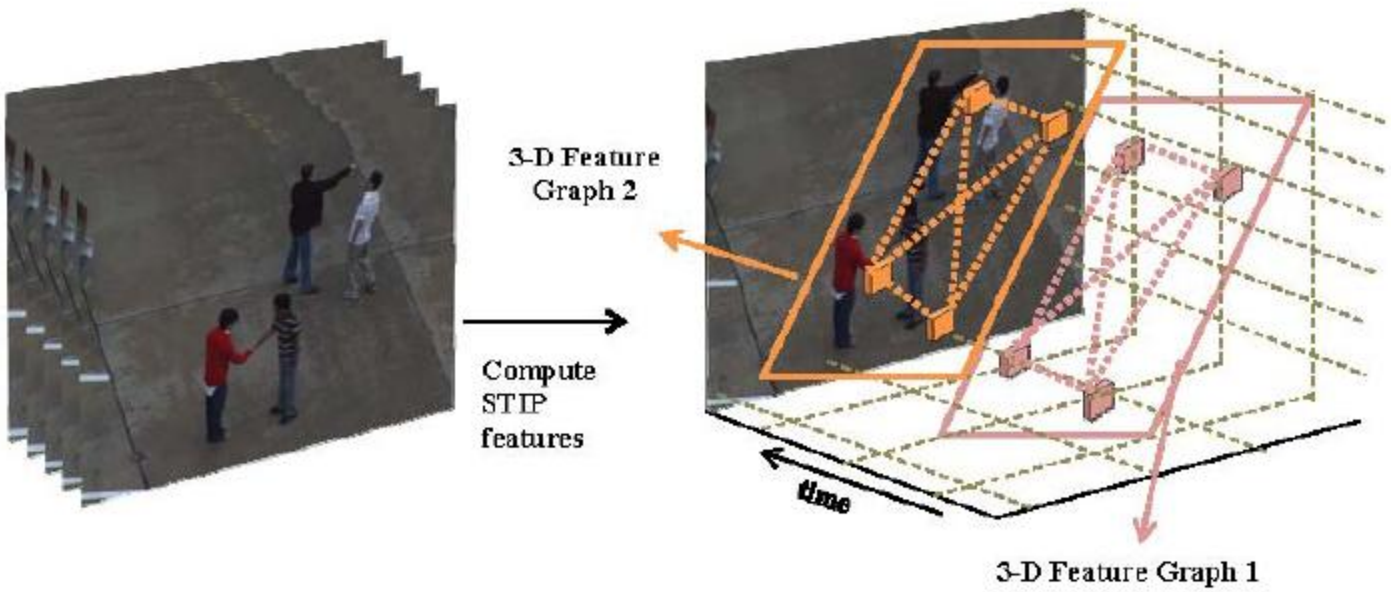
	answerPhone	chopBanana	dialPhone	drinkWater	eatBanana	eatSnack	lookupInPhonebook	peelBanana	useSilverware	writeOnWhiteboard
answerPhone	.47		.40	.07	.07					
chopBanana	.07	.73		.13			.07			
dialPhone	.33		.40		.20		.07			
drinkWater		.07		.53				.07	.33	
eatBanana	.07		.13		.67	.13				
eatSnack						.60	.07	.33		
lookupInPhonebook							1.0			
peelBanana						.33	.13	.47	.07	
useSilverware				.07			.07		.87	
writeOnWhiteboard										1.0

Trajectory of Local Features Result

Method	Percent Correct
Temporal Templates [6]	33
Spatio-Temporal Cuboids [7]	36
Space-Time Interest Points [12]	59
Velocity Histories (Sec. 3)	63
Latent Velocity Histories (Sec. 7)	67
Augmented Velocity Histories (Sec. 6)	89

String of Feature Graphs

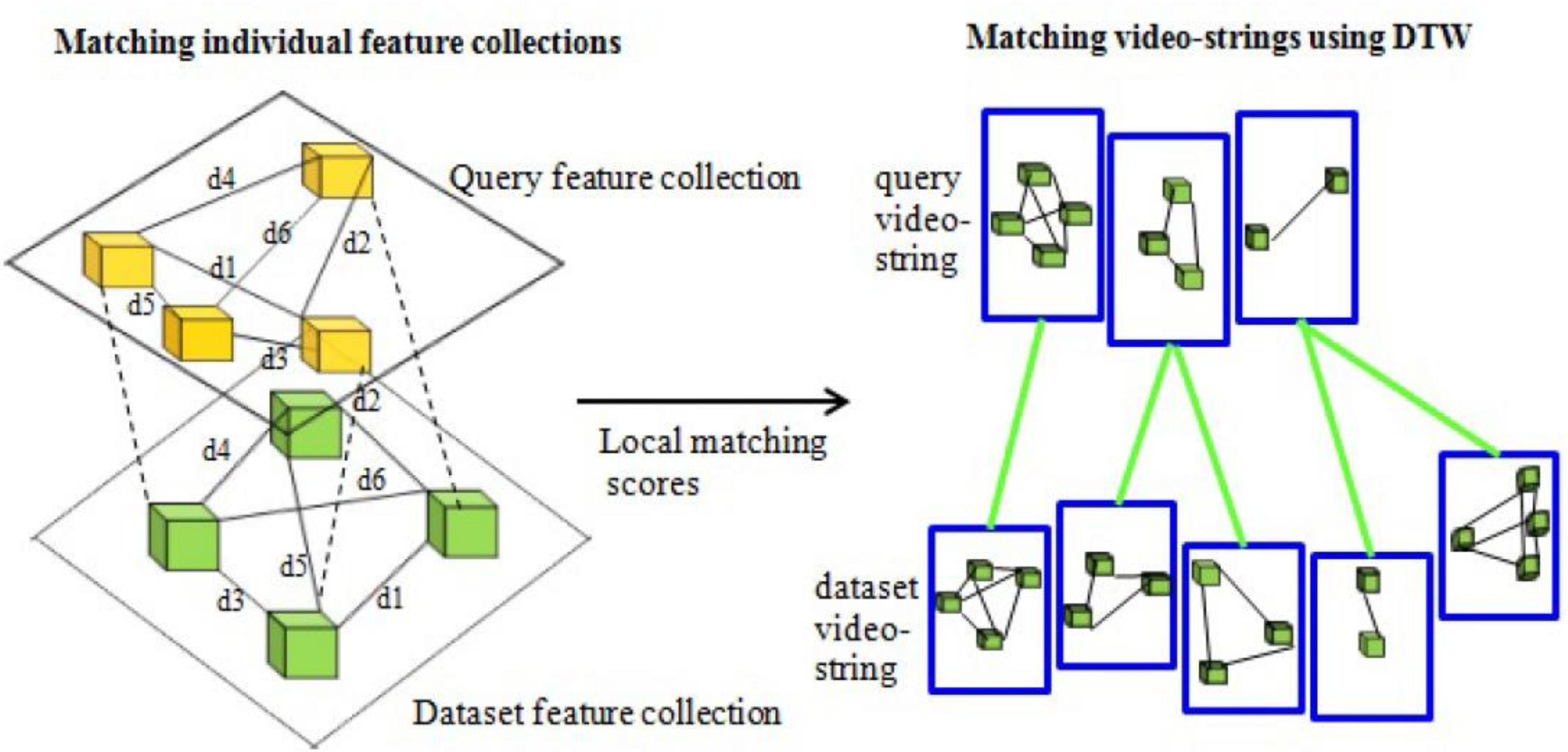
- [Gaur et al. ICCV 11]
- Interactive activities among different people



Match Feature Graphs

V1

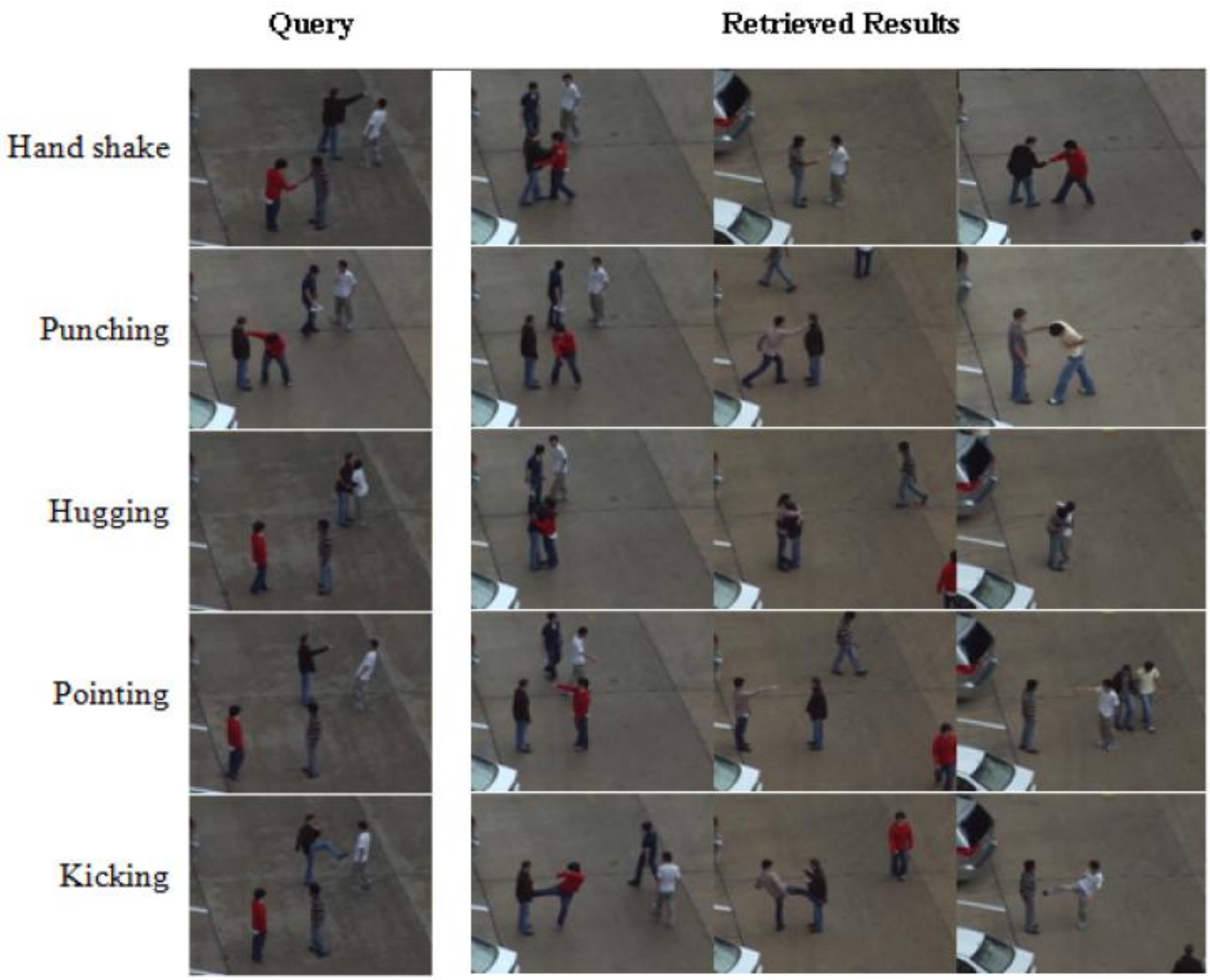
V2



Interactive Activities

Interactive Activities:
shake hands, hug, kick, point, push, punch

Example Results



Group-level activities

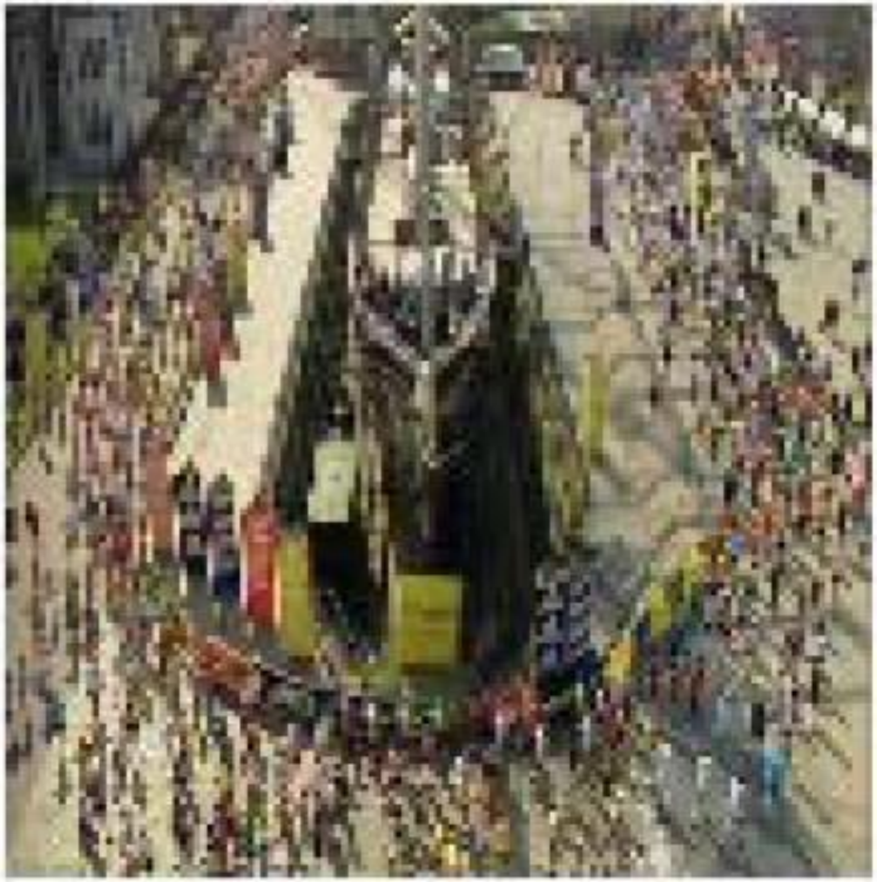
Conclusion

- Activity Recognition
  - Single Person
  - Multiple Persons
- Environments
  - Controlled environment, Stable cameras
  - Complex scenes – youtube, movie
- Approach
  - Holistic / Body part
  - Space-time local features
  - Incorporate long-range dependencies

Discussion

- Need for a large dataset of more activities
  - Current dataset: around 10 activity categories
  - ActivityNet?
  - A hierarchical dataset: high jump, long jump, ski jump
- Current algorithm is far from perfect
  - More suitable features?
- Speed is important
  - Avoid processing every frame

Events in Crowd



Images from [Wu et al. CVPR 2010]

Thank you