Evaluation of local spatio-temporal features for action recognition

Heng WANG[1,3], Muhammad Muneeb ULLAH[2], Alexander KLÄSER[1], Ivan LAPTEV[2], Cordelia SCHMID[1]

[1]LEAR, INRIA, LJK – Grenoble, France
[2]VISTA, INRIA – Rennes, France
[3]LIAMA, NLPR, CASIA – Beijing, China

Problem statement

- Local space-time features have become popular for action recognition in videos
- Several methods exist for *detection* and *description* of local spatio-temporal feature
- Existing comparisons are limited [Laptev'04, Dollar'05, Scovanner'07, Jhuang'07, Kläser'08, Laptev'08, Willems'08]
  - Different experimental settings
  - Different datasets
  - Evaluations limited to only few descriptors

Goal of this work

- Provide a common evaluation setup
  - Same datasets (varying difficulty): KTH, UCF sports, Hollywood2
  - Same train / test data
  - Same classification method
- Carry out a systematic evaluation of detector-descriptor combinations

Outline

- Action recognition framework
- Feature detectors
- Feature descriptors
- Experimental results

Action recognition framework

Feature detectors

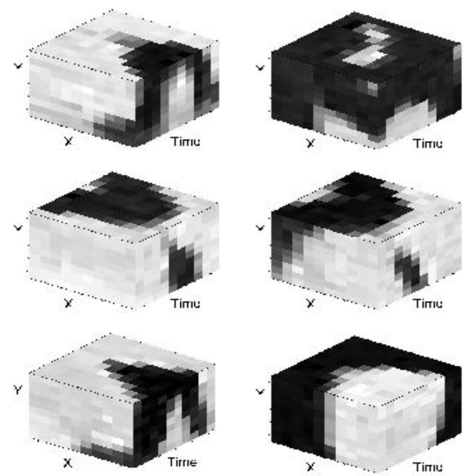
Feature descriptors

Experimental results

Detection + description of features

Detection of feature / interest points

Space-time patches



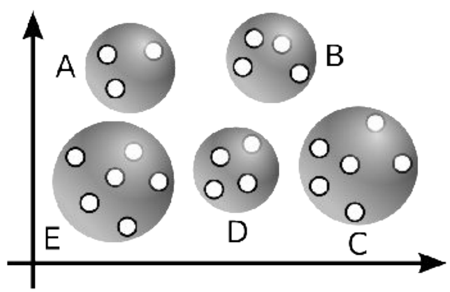
Description of space-time patches

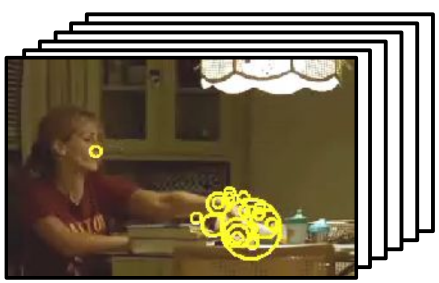
Patch representation as feature vector $v = (v_1, v_2, ..., v_n)$

Bag-of-words representation

Bag of space-time features + SVM [Schuldt'04, Niebles'06, Zhang'07]

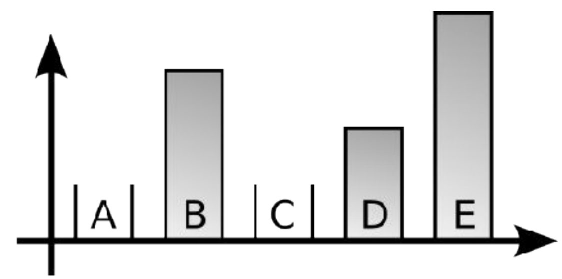
Training feature vectors are clustered with k-means (k=4000)





Each feature vector is assigned to its closest cluster center (visual word)

Classification with non-linear SVM and χ^2-kernel



An entire video sequence is represented as occurrence histogram of visual words

Action recognition framework

Feature detectors

Feature descriptors

Experimental results

Spatio-temporal feature detectors

Evaluation of 4 types of feature detectors

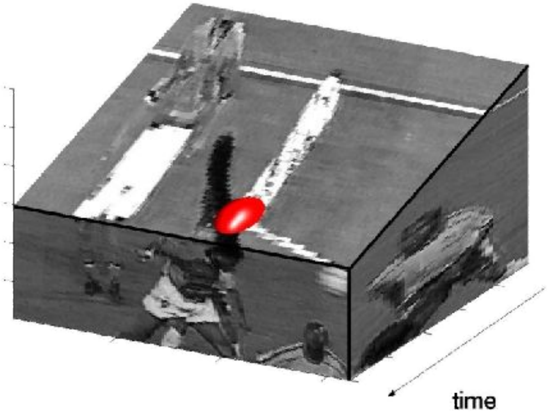
- Harris3D [Laptev'05]
- Cuboid [Dollar'05]
- Hessian [Willems'08]
- Dense

Harris3D detector [Laptev'05]

- Space-time corner detector

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

$$\mu = \begin{pmatrix} I_x I_x & I_x I_y & I_x I_t \\ I_x I_y & I_y I_y & I_y I_t \\ I_x I_t & I_y I_t & I_t I_t \end{pmatrix} * g(\cdot;\sigma,\tau)$$



- Any spatial and temporal corner is detected
- Dense scale sampling (no explicit scale selection)

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

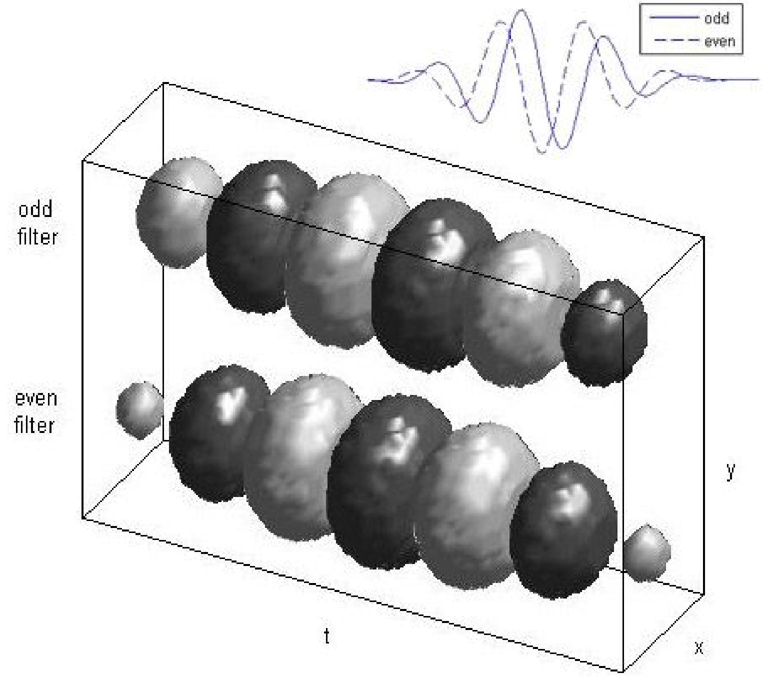
Cuboid detector [Dollar'05]

- Space-time detector based on temporal Gabor filters
- Response function:

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

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

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



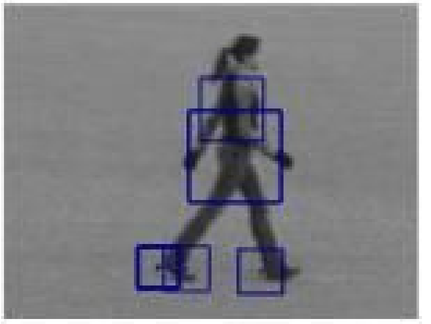
- Detects regions with spatially distinguishing characteristics undergoing a complex motion

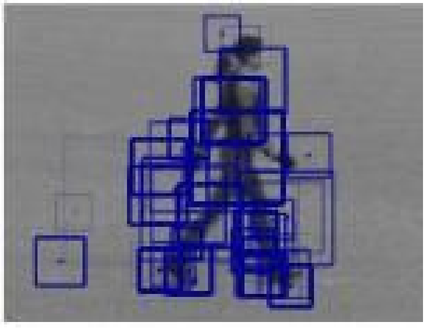
Hessian detector [Willems'08]

- Spatio-temporal extension of the Hessian saliency measure [Lindberg'98]
- Strength of interest point computed with the determinant of the Hessian matrix:

$$H(\cdot;\sigma^2,\tau^2) = \begin{pmatrix} L_{xx} & L_{xy} & L_{xt} \\ L_{yx} & L_{yy} & L_{yt} \\ L_{tx} & L_{ty} & L_{tt} \end{pmatrix}$$

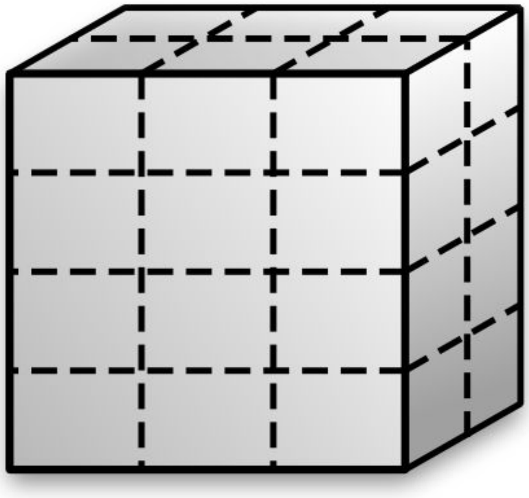
$$S = |det(H)|$$





- Approximation with integral videos
- Detects spatio-temporal 'blobs'

Dense Sampling



- Motivation: dense sampling outperforms interest points in object recognition
  [Fei-Fei'05, Jurie'05]
- For videos: extract 3D patches at regular positions (x, y, t) with varying scales (sigma, tau)
- Spatial and temporal overlap of 50%
- Minimum size: 18x18x10, scale factor: sqrt(2)

Illustration of detectors

Harris

Cuboid

Hessian

Action recognition framework

Feature detectors

Feature descriptors

Experimental results

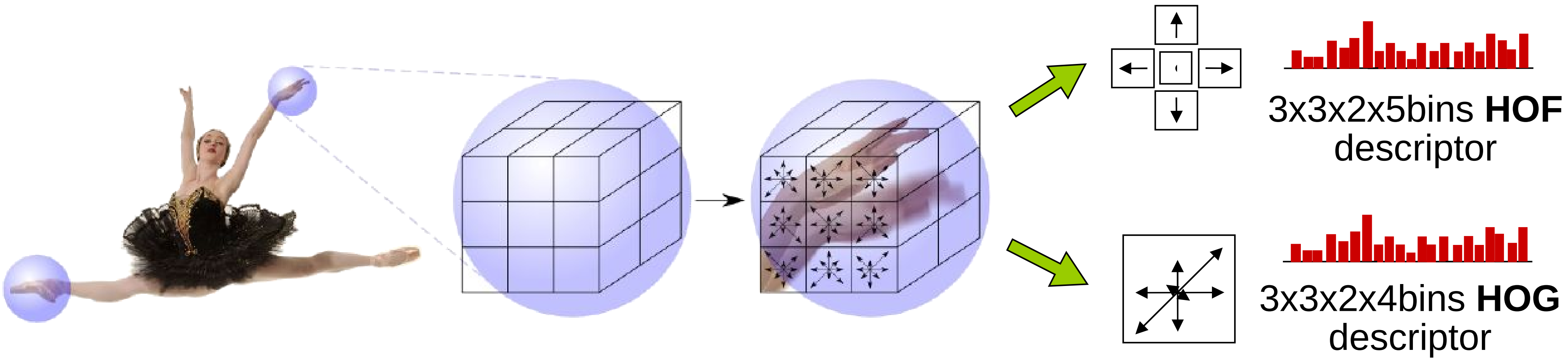
Spatio-temporal feature descriptors

Evaluation of 4 types of feature descriptors

- HOG/HOF [Laptev'08]
- Cuboid [Dollar'05]
- HOG3D [Kläser'08]
- Extended SURF [Willems'08]

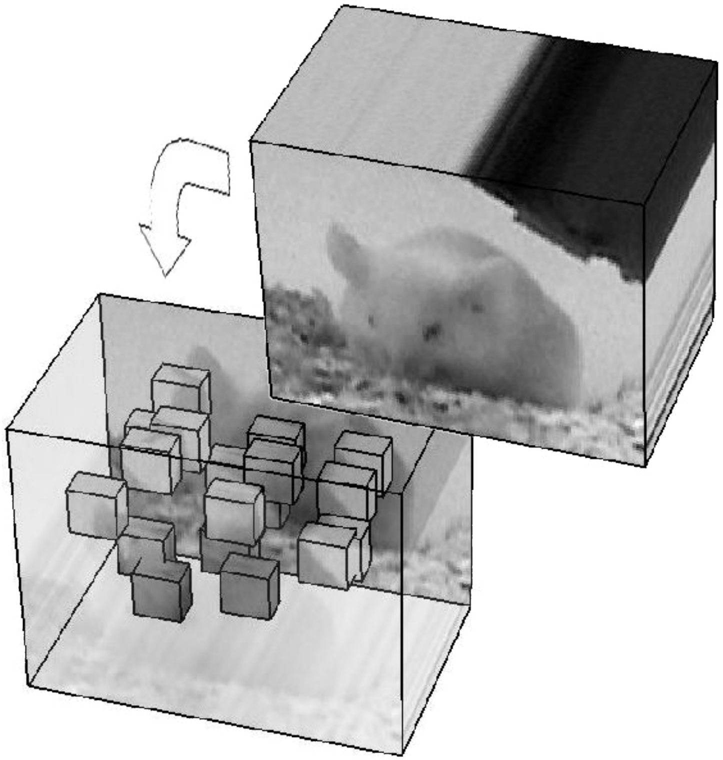
HOG/HOF descriptor [Laptev'08]

- Based on histograms of oriented (spatial) gradients (HOG) + histogram of optical flow (HOF)
- 3D patch is divided into a grid of cells
- Each cell is described with HOG/HOF



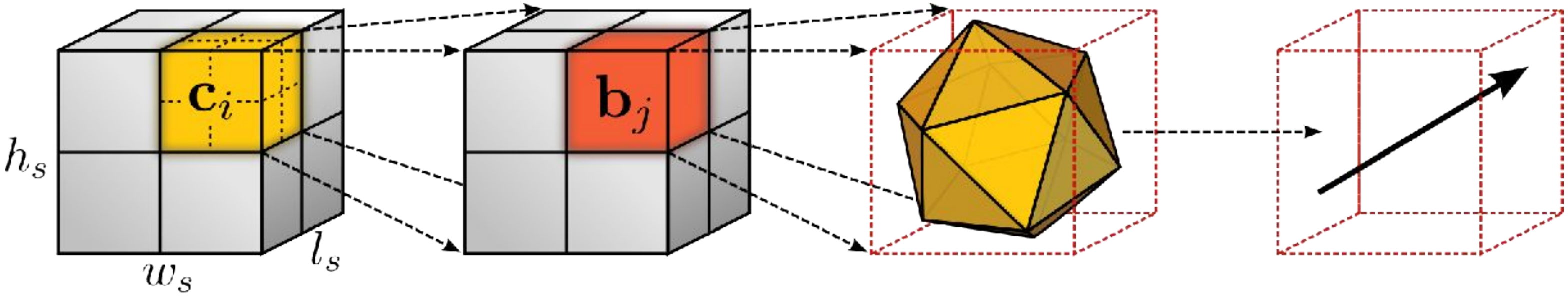
Cuboid descriptor [Dollar'05]

- 3D patch is described by its gradient values
- Gradient values for each pixel are concatenated
- PCA reduces dimensionality



HOG3D descriptor [Kläser'08]

- An extension of SIFT descriptor to videos
- Based on histograms of 3D gradient orientations
- Uniform quantization via regular polyhedrons
- Combines shape and motion information



(a) full descriptor (with 2x2x2 histogram cells) **(b)** histogram computation (over 2x2x2 sub-blocks) **(c)** gradient orientation quantization **(d)** mean gradient computation

E-SURF descriptor [Willems'08]

- E-SURF: an extension of SURF descriptor [Bay'06] to videos
- 3D cuboid is divided into cells
- Bins are filled with weighted sums of responses of the axis-aligned Haar-wavelets dx, dy, dt

$$\boldsymbol{v} = (\sum d_x, \sum d_y, \sum d_t)$$

Action recognition framework

Feature detectors

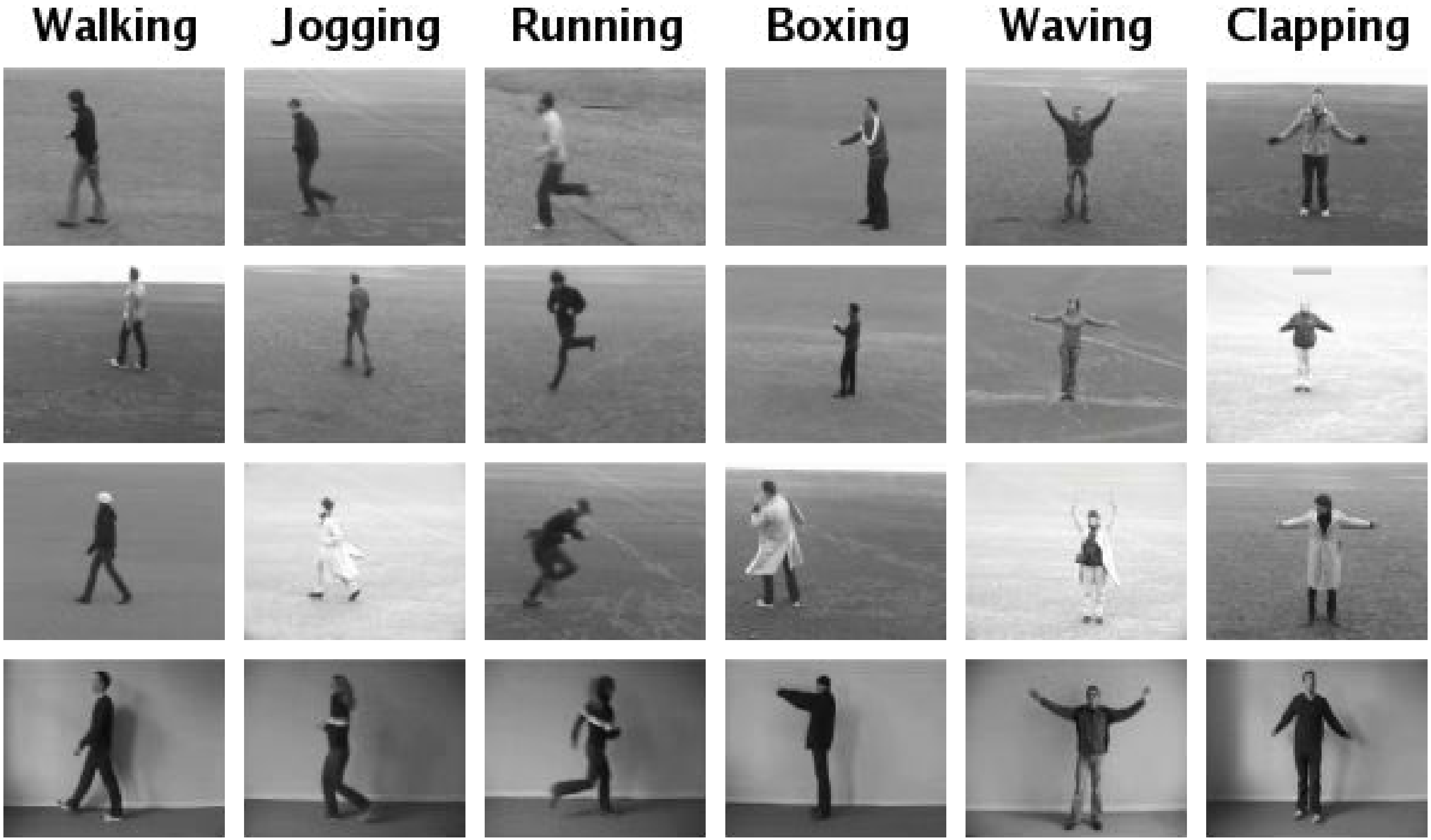
Feature descriptors

Experimental results

Dataset: KTH actions

- 10 action classes
- 25 people performing in 4 different scenarios
  - Training samples from 16 people
  - Testing samples from 9 people
- In total 2391 video samples
- Note: homogenous and static background
- Measure: *average accuracy* over all classes
- State-of-the-art: 91.8% [Laptev'08]

KTH actions – samples



KTH actions – results

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

- Best results for Harris3D + HOF
- Good results for Harris3D & Cuboids detector and HOG/HOF & HOG3D descriptor
- Dense features worse than interest points
  - Large number of features on static background

Dataset: UCF sports

- 10 different (sports) action classes
- 150 video samples in total
  - We extend the dataset by flipping videos
- Evaluation via leave-one-out
- Measure: *average accuracy* over all classes
- State-of-the-art: 69.2% [Rodriguez'08]

UCF sports – samples

Swinging | Diving | Kicking | Lifting | Horse Riding

Running | Skateboard | High-bar | Golf | Walking

UCF sports – results

Detectors

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

- Best results for Dense + HOG3D
- Good results for Dense and HOG/HOF
- Cuboids detector: performs well with HOG3D

Dataset: Hollywood2 actions

- 12 different action classes
- In total from 69 different Hollywood movies
- 1707 video samples in total
- Separate movies for training / testing
- Measure: *mean average precision* over all classes

Hollywood2 actions – samples

Answer phone

Drive car

Eat

Fight

Get out of car

Hand shake

Hug

Kiss

Run

Sit down

Sit up

Stand up

Hollywood2 actions – results

Detectors

Descriptors

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

- Best results for Dense + HOG/HOF
- Good results for HOG/HOF

Conclusion

- Dense sampling consistently outperforms all the tested detectors in realistic settings (UCF + Hollywood2)
  - Importance of realistic video data
  - Limitations of current feature detectors
  - Note: large number of features (15-20 times more)
- Detectors: Harris3D, Cuboids, and Hessian provide overall similar results (interest points better than Dense on KTH)
- Descriptors overall ranking:
  - HOG/HOF > HOG3D > Cuboids > ESURF & HOG
  - Combination of gradients + optical flow seems good choice
- This is the first step... we need to go further...

Do you have questions?

Computational complexity

	Harris3D + HOG/HOF	Hessian + ESURF	Cuboid Det.+Desc.	Dense + HOG3D	Dense + HOG/HOF
Frames/sec	1.6	4.6	0.9	0.8	1.2
Features/frame	31	19	44	643	643

- Dollar extracts the most dense features and is the slowest (0.9 FPS)
- Hessian extracts the most sparse features and is the fastest (4.6 FPS)
- Dense sampling extracts many more features compared to interest point detectors