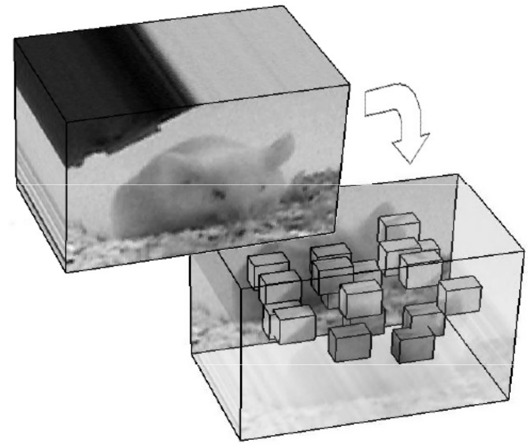
Keypoint-Based Action Recognition

Presenter: Jianchao Yang

Course Instructor: Prof. Derek Hoiem

Papers to discuss

- *Behavior recognition via sparse spatio-temporal features.*



- *Learning realistic human actions from movies.*

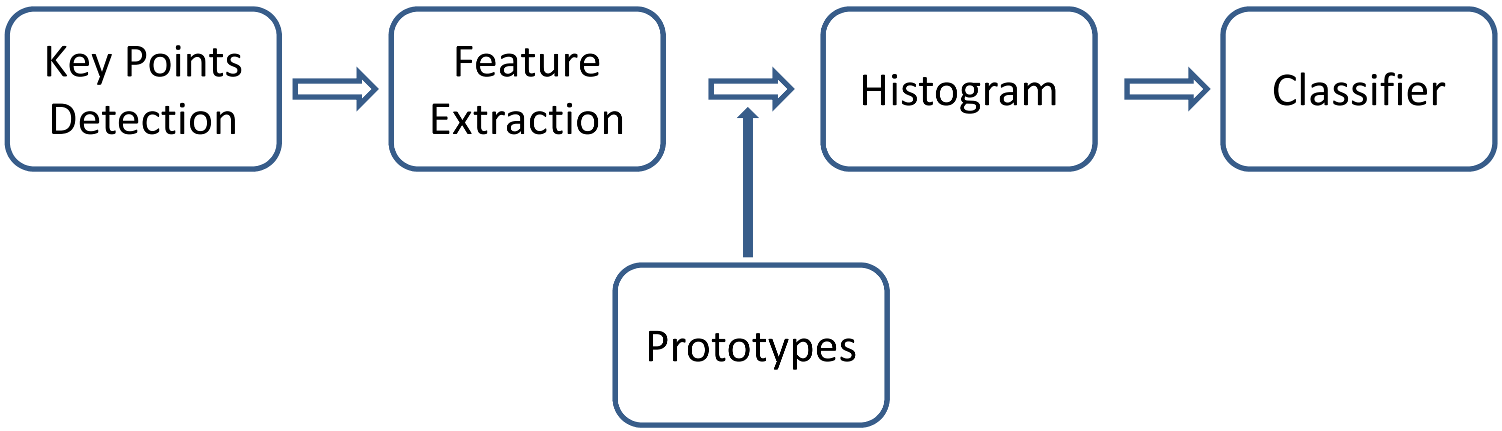
Behavior Recognition via Sparse Spatio-Temporal Features

- Motivated by the success application of key points in object recognition

- **Designed a spatio-temporal feature for behavior recognition**

Approach

- Similar to what seen in object recognition

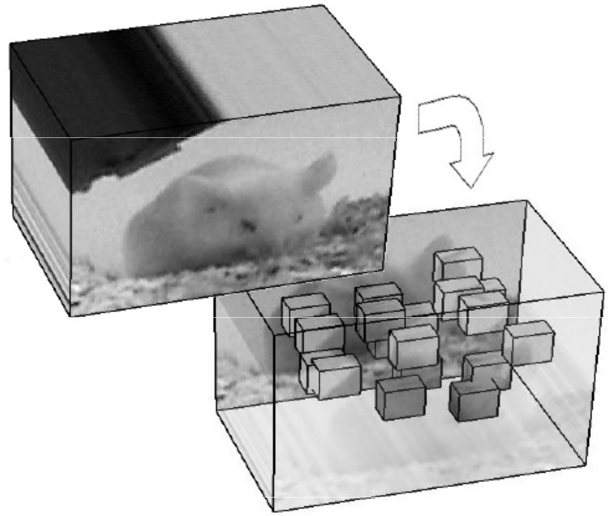


Keypoints detection

- Extension from 2D

- Localization proceeds along the spatial dimensions **x** and **y**, as well as the temporal dimension **t**.

- 3D corners too rare



Keypoints detection (cont')

- Response function:

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

  - Spatial kernel $g(x, y; \sigma)$ is 2D Gaussian
  - Temporal kernel

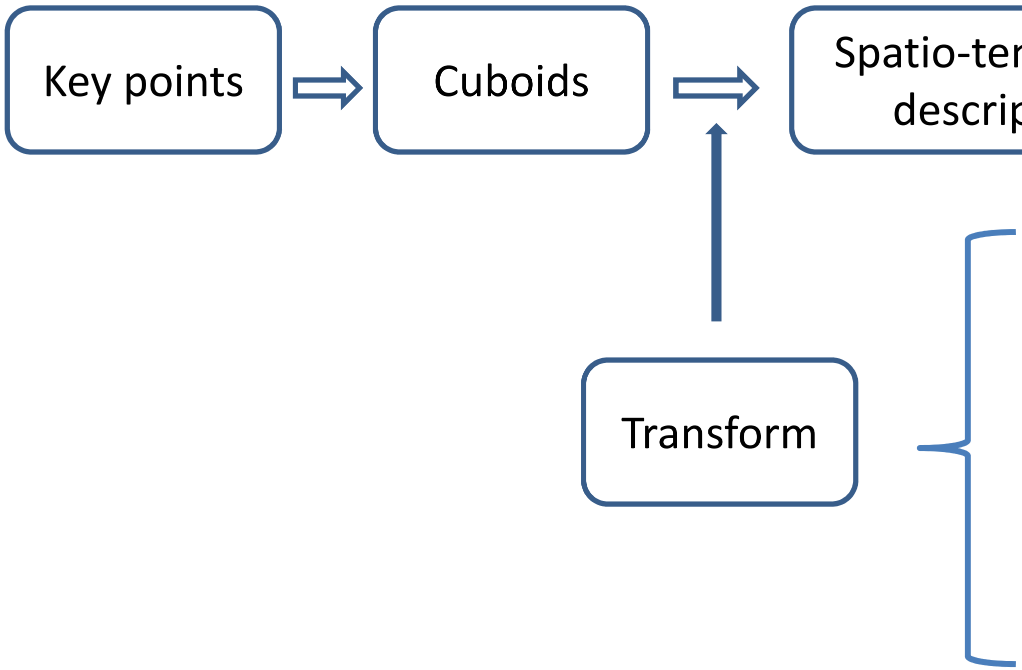
$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$$

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

Keypoints detection (cont')

- Keypoints
  - By pooling maxima of the filter responses
  - Emphasize **temporal** information other than spatial information
  - Strong response to periodic motions
  - Does not respond to pure translation motion
  - Totally unsupervised

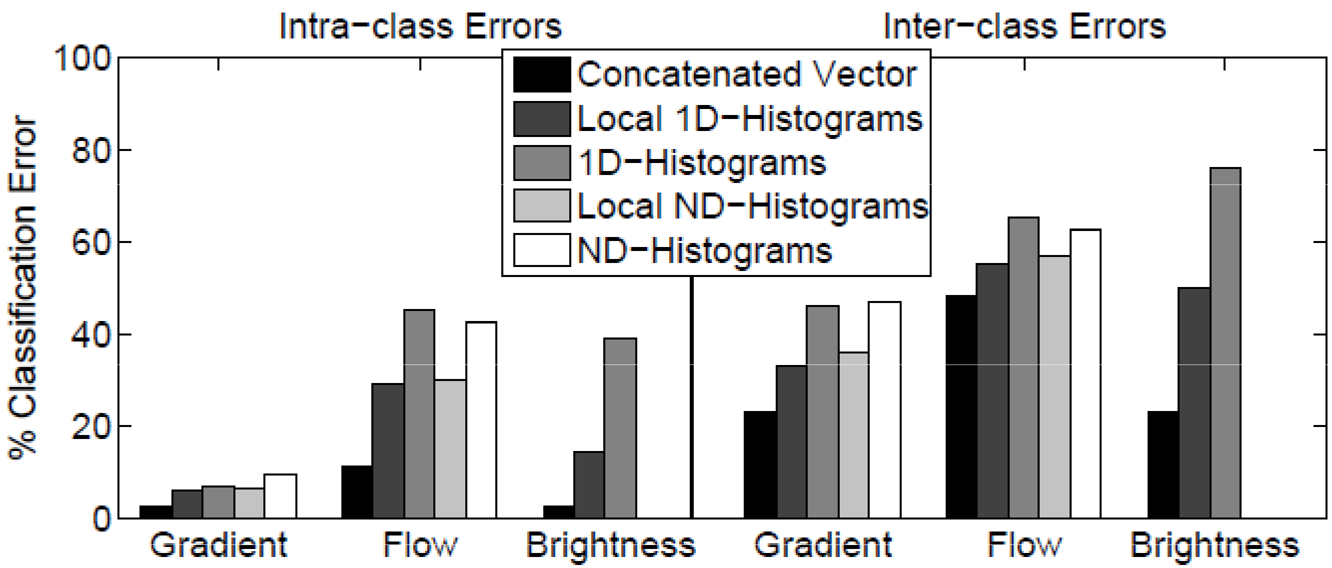
Cuboid descriptor



Descriptor:
Normalized pixel values;
Gradients;
Windowed optical flow, etc.

Transform:
Vectorize directly;
Histogram (global or local).

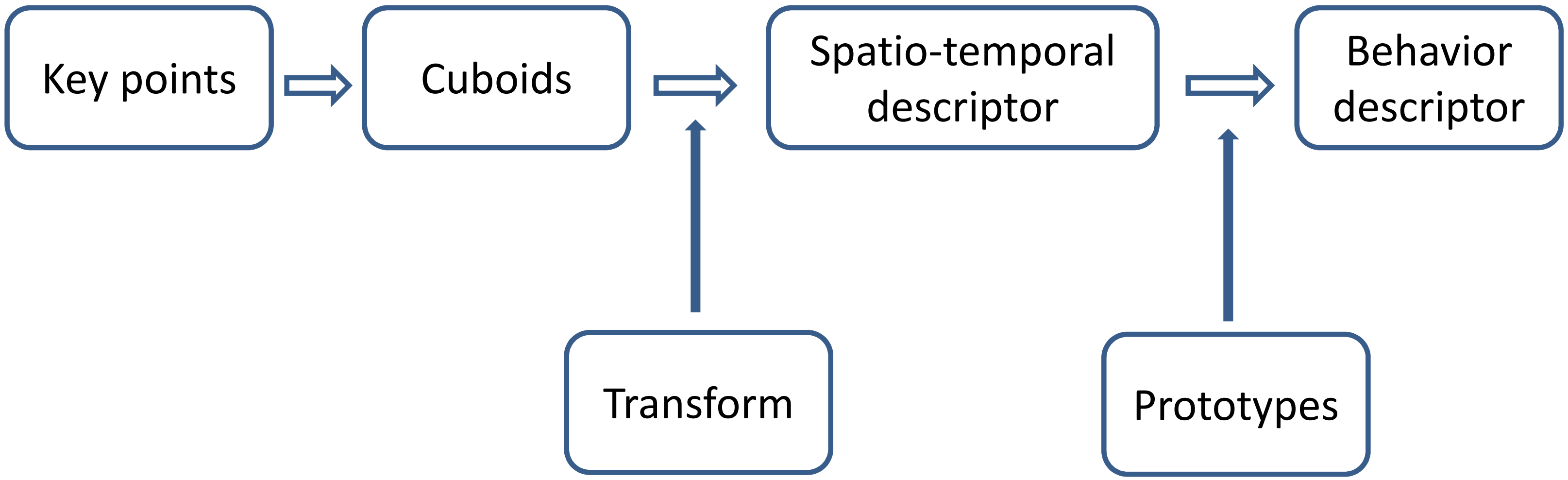
Cuboid descriptor (cont')



Gradient is best!

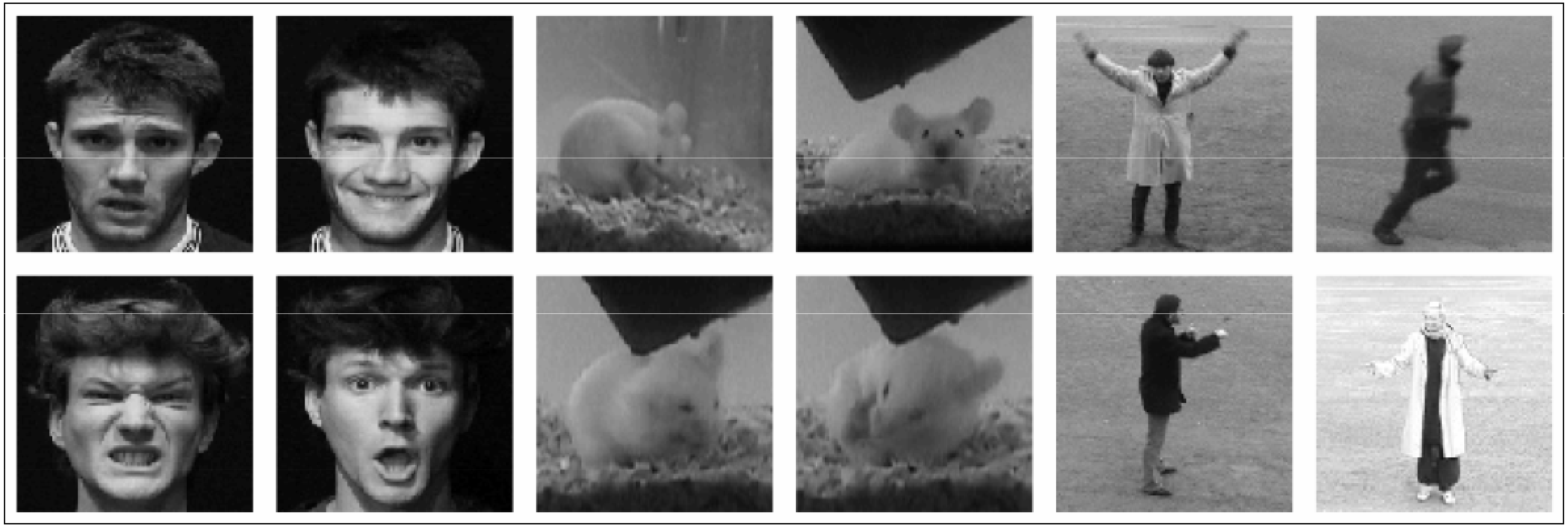
Vectorize directly is best! ??

Behavior descriptor

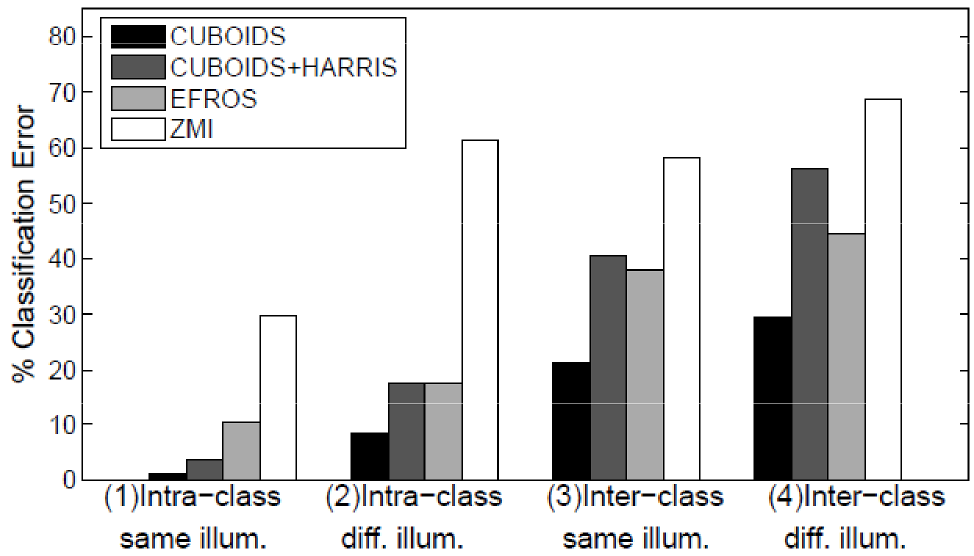


Experiment results

- Datasets: facial expression, mouse, human actions

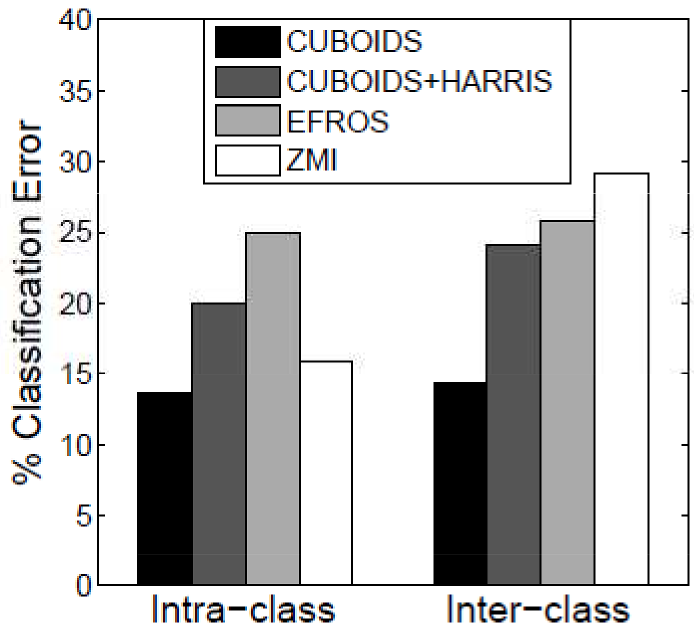


Experiments results (cont')



Human Facial Expression Database.

Mouse Database.



Learning realistic human actions from movies

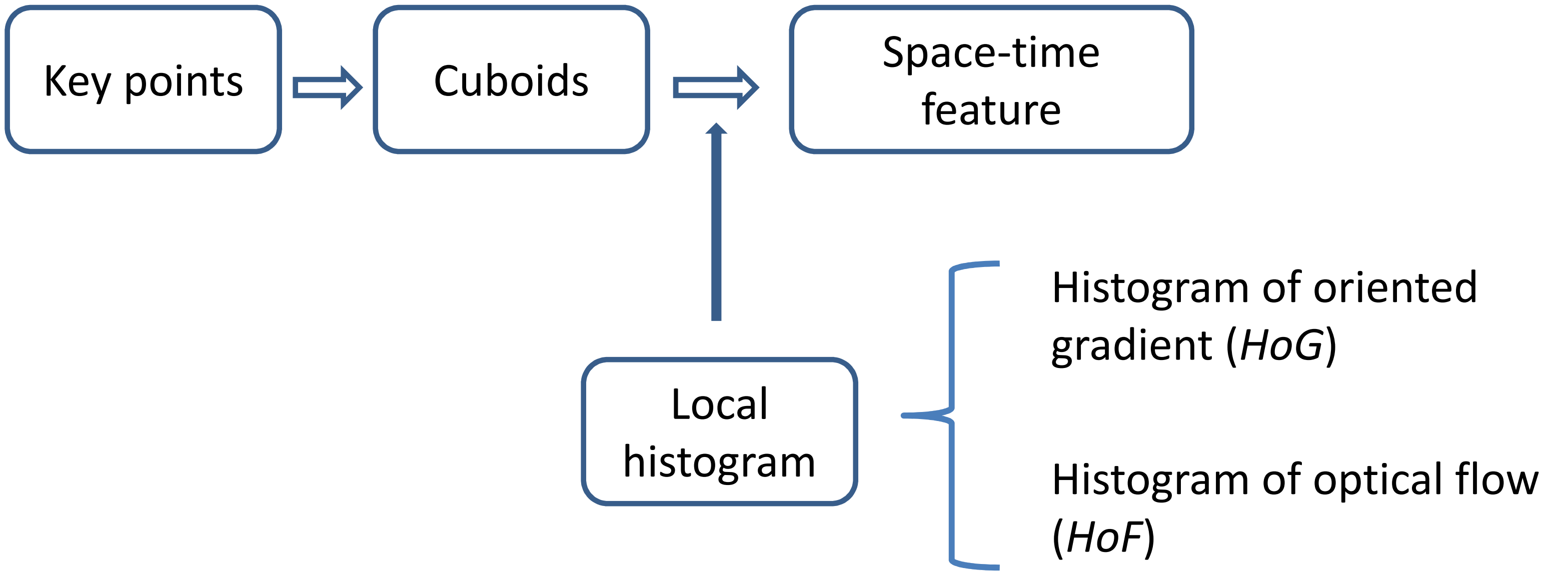
- Automatic annotation of human actions in video.
- **Video classification by space-time features.**

Bag-of-feature approach

- Extension of recent advances in bag-of-feature approaches
  - Spatial pyramid → more general spatial grids
  - Fixed weights for each pyramid level → optimized
  - Spatial grid → space-time grids

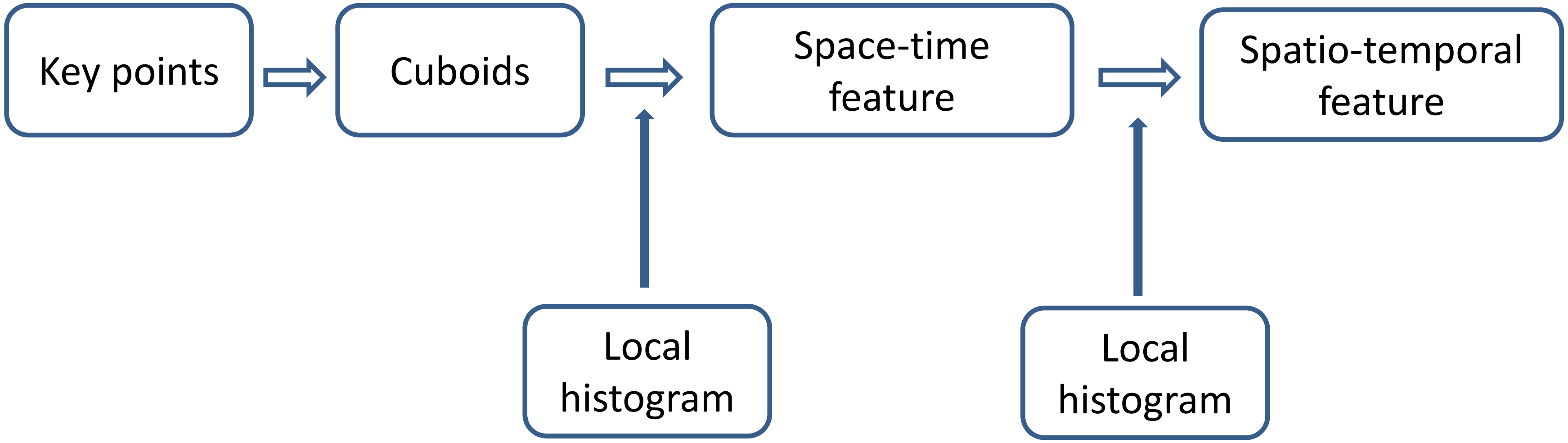
Space-time features

- Interest point detection: Harris operator

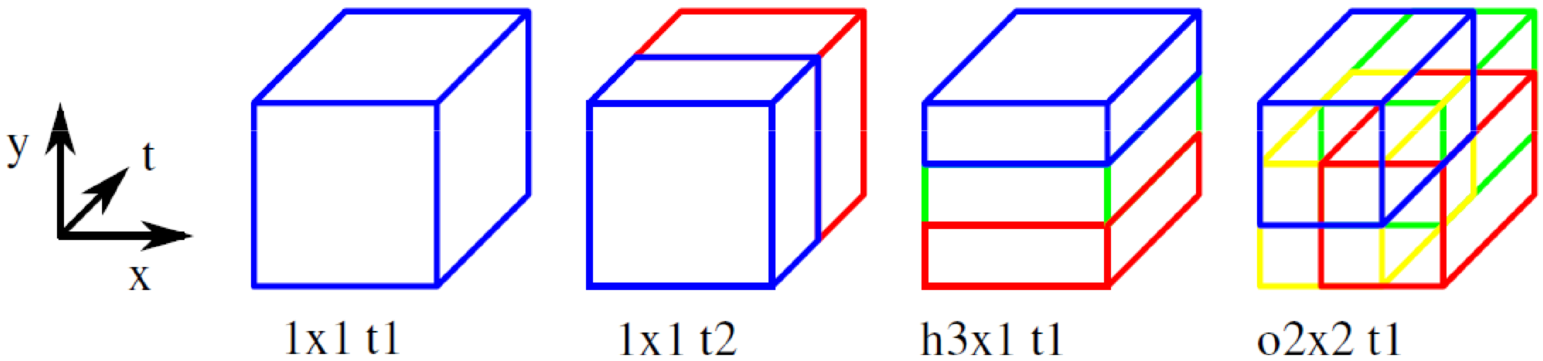


Spatio-temporal bag-of-features

- Hierarchical structure

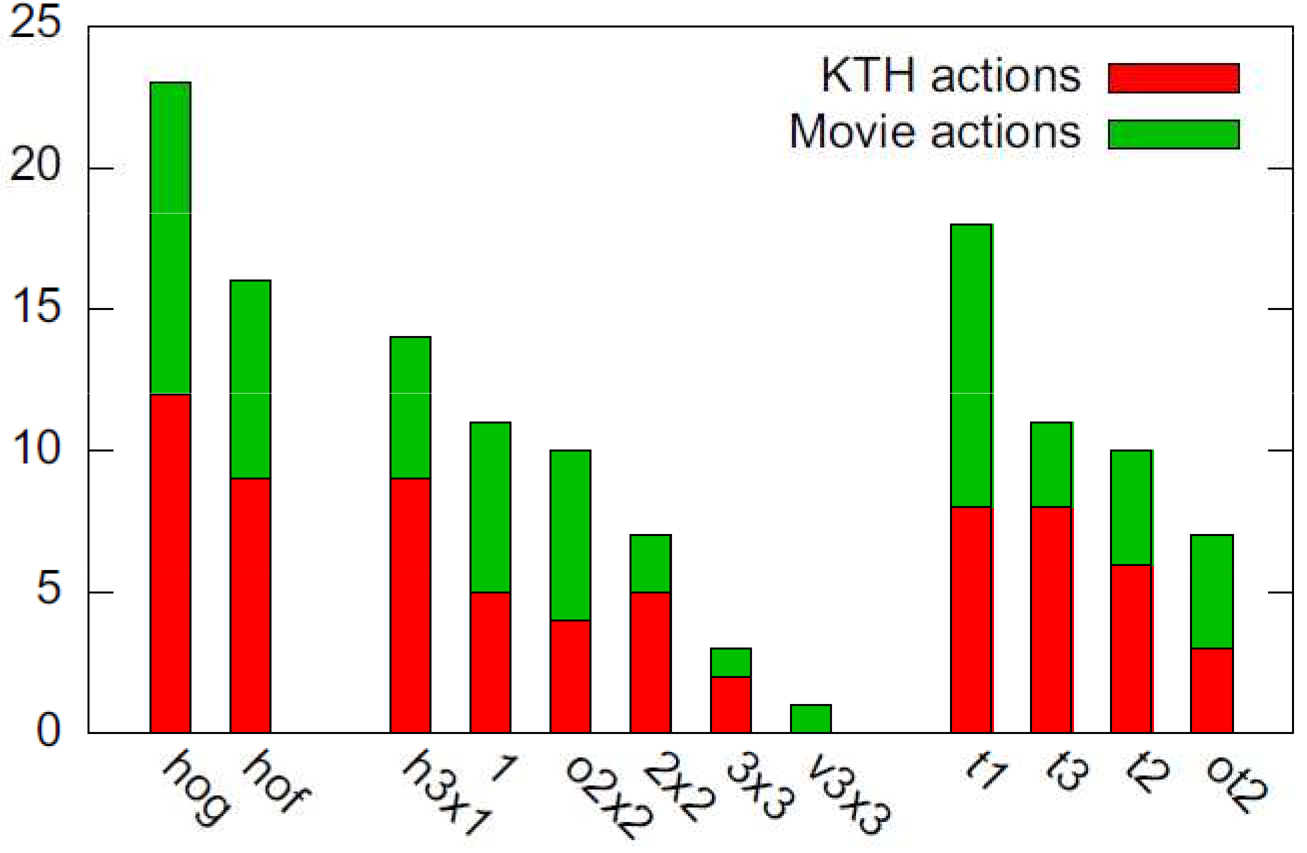


Spatio-temporal grids



Experiment results

- Evaluation of spatio-temporal grids

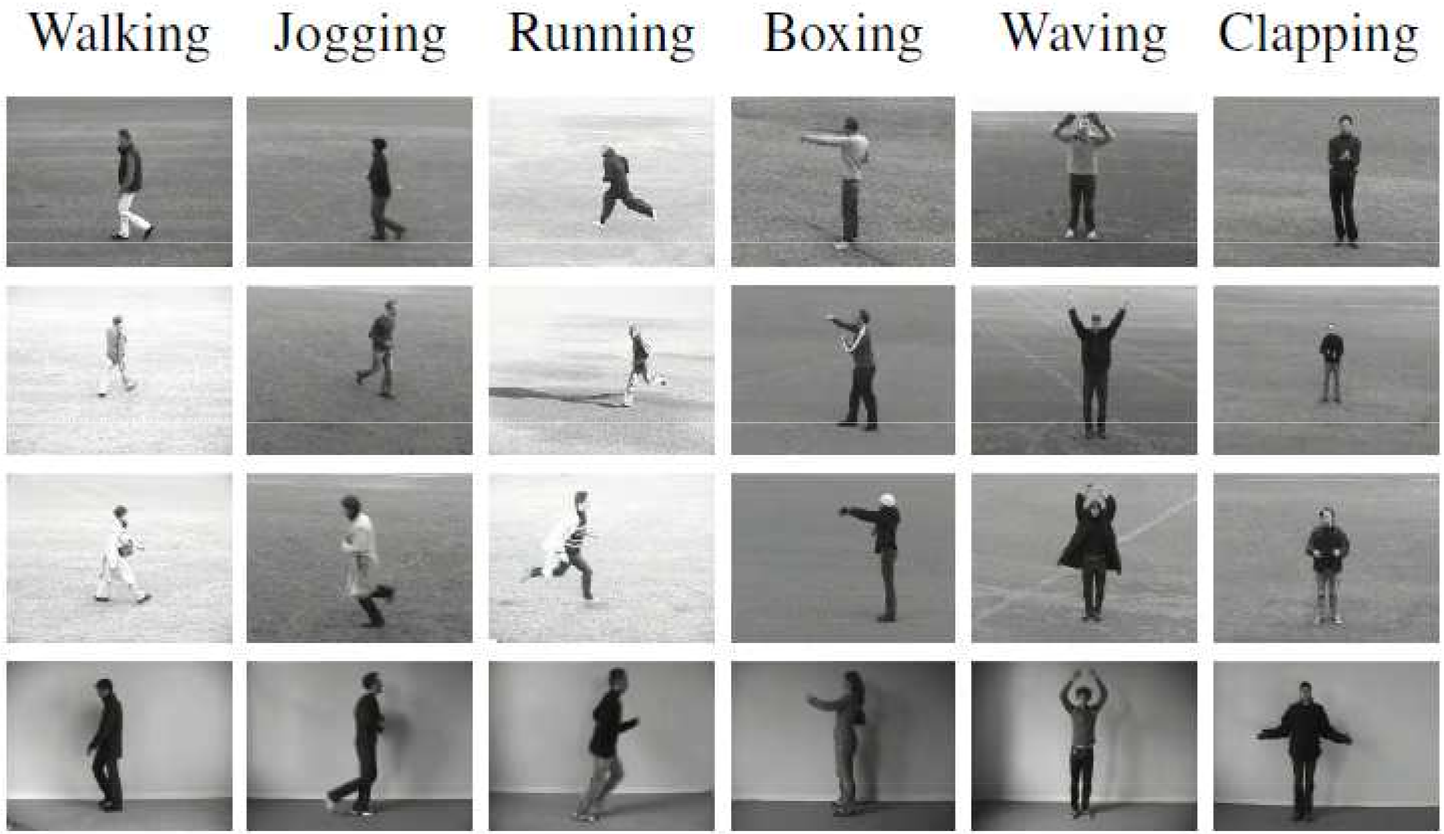


Experiments results (cont')

Task	HoG BoF	HoF BoF	Best channel	Best combination
KTH multi-class	81.6%	89.7%	91.1% (hof h3x1 t3)	91.8% (hof 1 t2, hog 1 t3)
Action AnswerPhone	13.4%	24.6%	26.7% (hof h3x1 t3)	32.1% (hof o2x2 t1, hof h3x1 t3)
Action GetOutCar	21.9%	14.9%	22.5% (hof o2x2 1)	41.5% (hof o2x2 t1, hog h3x1 t1)
Action HandShake	18.6%	12.1%	23.7% (hog h3x1 1)	32.3% (hog h3x1 t1, hog o2x2 t3)
Action HugPerson	29.1%	17.4%	34.9% (hog h3x1 t2)	40.6% (hog 1 t2, hog o2x2 t2, hog h3x1 t2)
Action Kiss	52.0%	36.5%	52.0% (hog 1 1)	53.3% (hog 1 t1, hof 1 t1, hof o2x2 t1)
Action SitDown	29.1%	20.7%	37.8% (hog 1 t2)	38.6% (hog 1 t2, hog 1 t3)
Action SitUp	6.5%	5.7%	15.2% (hog h3x1 t2)	18.2% (hog o2x2 t1, hog o2x2 t2, hog h3x1 t2)
Action StandUp	45.4%	40.0%	45.4% (hog 1 1)	50.5% (hog 1 t1, hof 1 t2)

Experiment results (cont')

- KTH action database



Experiment results (cont')

	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%

Comments

- The two methods are extensions of key-points based image classification. Will dense descriptors be better?

- Key-points based methods work surprisingly well for image and sequence classification, why?

- Issues needed to address:
  - Discriminative key-points learning or design for the given task
  - Discriminative key-points selection for the given task
  - More efficient way to use location information