Event, Action & Activity Recognition - II

Rainer Stiefelhagen
Ziad Al-Halah (ziad.al-halah@kit.edu

27.01.2014

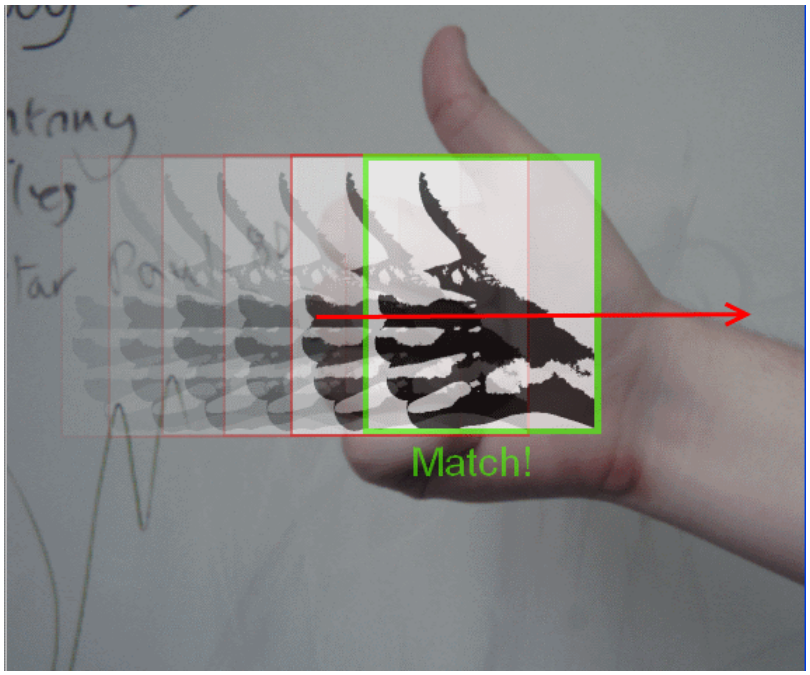
Overview

- Last lecture
  - Event recognition using time series classification methods
    - Left-to-right HMMs
    - Layered HMMs
- Today
  - Event recognition approaches inspired by object recognition systems
    - Template matching
    - Boosting
    - Bag-Of-Words

Example approaches

Template matching

Template Matching in Object Detection



- Basic concept:
  - Move template across the image and calculate similarity (eg. cross-correlation) with image patch
  - Detection, if similarity value is above a threshold

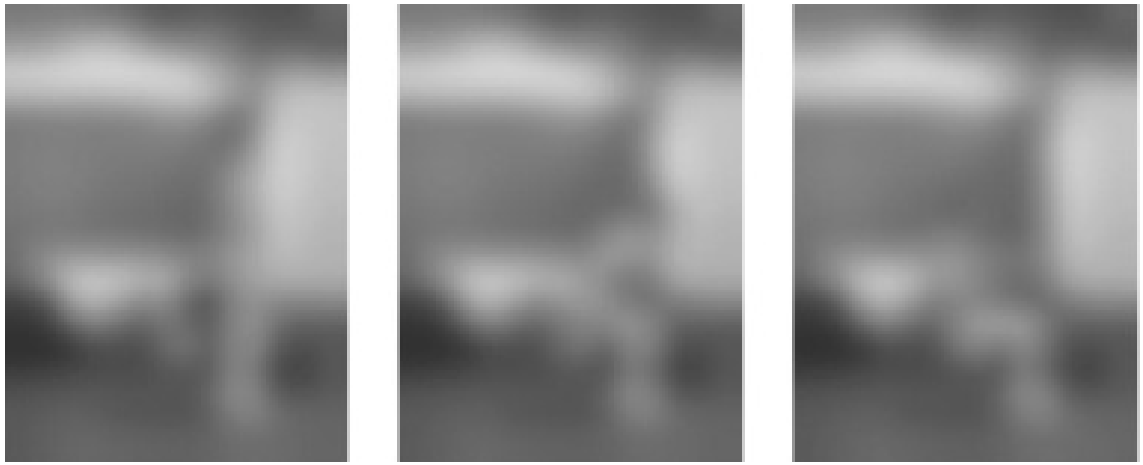
Recognition of Human Movement using Temporal Templates [Bobick01]





- Objective: Classify a set activities based on a person's motion
- Input:
  - Several close-up camera views
  - Static, indoor scene

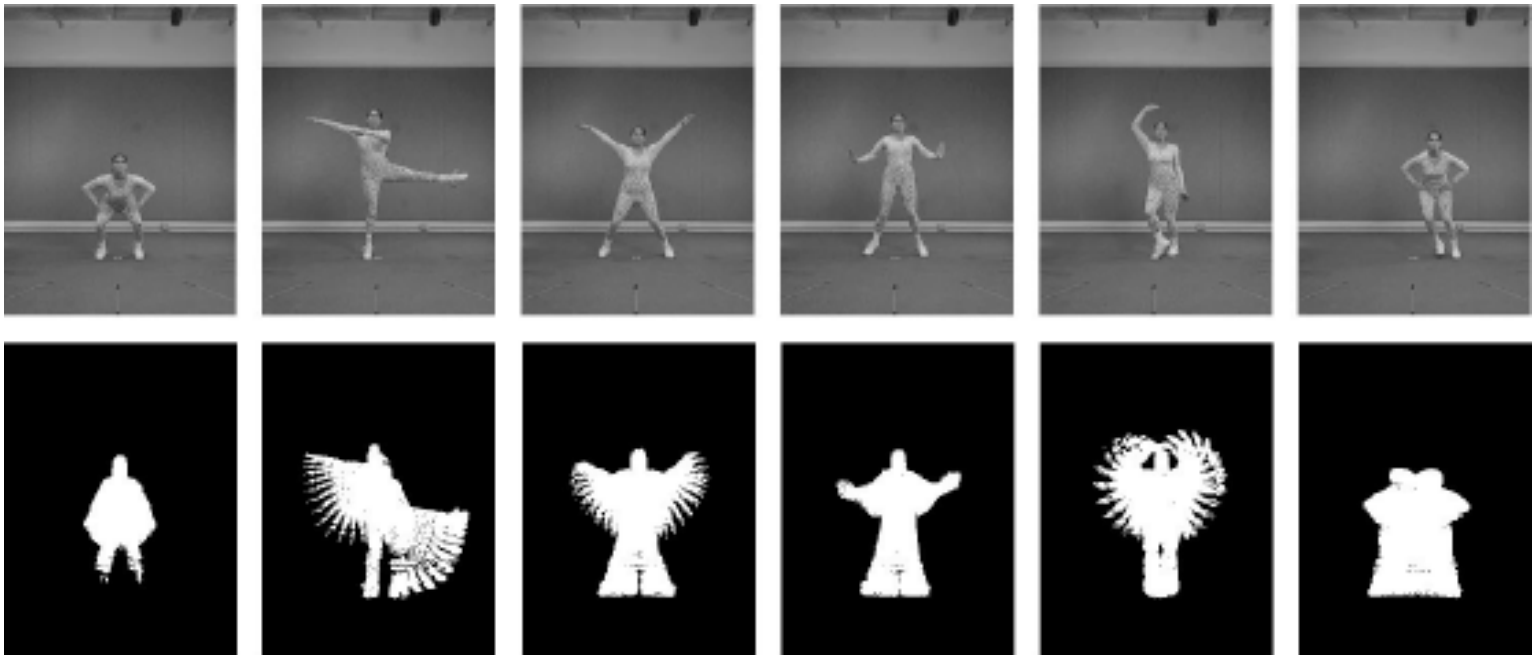
Motion



- Even with almost no structure in the video, humans can recognize activity through motion (walking, sitting down)
- No 2D/3D reconstruction of body model necessary
- Need to know:
  - Where is motion?
  - How is it moving?

Motion Features

- Motion Energy Image (MEI)



- Captures the information: Where is motion

MEI

- Let *I(x,y,t)* be an image sequence
- Let *D(x,y,t)* be a binary image sequence indicating regions of motion (e.g. difference image)

- The MEI is defined as:

$$E_\tau(x,y,t) = \bigcup_{i=0}^{\tau-1} D(x,y,t-i)$$

- τ is the size of the observation window (1-2 secs)

Motion Features

- Motion History Image (MHI)
- Captures the information: How is motion done

sit-down

sit-down MHI

arms-wave

arms-wave MHI

crouch-down

crouch-down MHI

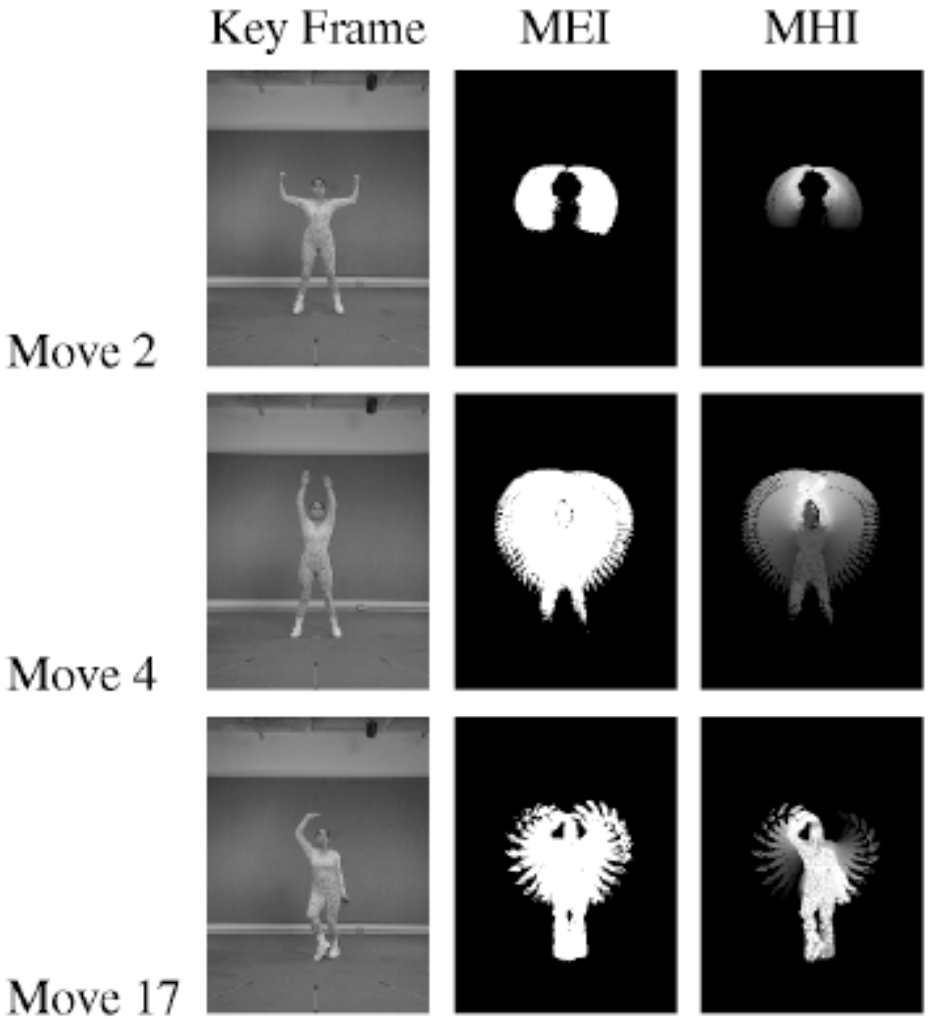
MHI

$$H_\tau(x,y,t) = \begin{cases} \tau & \text{if } D(x,y,t) = 1 \\ \max\left(0, H_\tau(x,y,t-1) - 1\right) & \text{otherwise.} \end{cases}$$

- Result: More recently moving pixels are brighter
- Note: MEI can be generated by thresholding the MHI above zero

Why both MEI and MHI?

- For some moves, MEIs are similar, for others, MHIs are similar
- MEI and MHI capture different characteristics of motion
  - "where" and "how"



Matching Temporal Templates

- Training
  - Collect training examples for each move from a variety of viewing angles
  - Generate MEIs and MHIs
  - Calculate scale and orientation invariant features (Hu-moments) on images → Feature vector
  - Build statistical model of the moments (mean m_j, covariance matrix Σ_j)
- Recognition:
  - Calculate Mahalanobis distance between moment description of input and each of the stored movements

$$\Delta_j(x) = (x - m_j)^T \Sigma_j^{-1} (x - m_j)$$

Moments as Shape Descriptors

- Idea: a density distribution (e.g. an image) is well described by its moments
  - → use statistical properties (moments) to describe their shape

- Two-dimensional (p+q)th order moments of a density distribution function:

$$m_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q \rho(x,y)\,dx\,dy,$$

In digital images:

$$M_{ij} = \sum_x \sum_y x^i y^j I(x,y)$$

- Central moments (invariant to translation):

$$\mu_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\bar{x})^p (y-\bar{y})^q \rho(x,y)\,d(x-\bar{x})\,d(y-\bar{y})$$

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q f(x,y)$$

  - where $\bar{x} = m_{10}/m_{00}$, $\bar{y} = m_{01}/m_{00}$ (centroids)

Hu-moments [Hu 1962]

- Goal: Find translation-, scale- and rotation-invariant moments to do pattern recognition

- Central moments (first four orders):

$$
\begin{aligned}
\mu_{00} &= m_{00} \equiv \mu \\
\mu_{10} &= 0 \\
\mu_{01} &= 0 \\
\mu_{20} &= m_{20} - \mu\bar{x}^2 \\
\mu_{11} &= m_{11} - \mu\bar{x}\bar{y} \\
\mu_{02} &= m_{02} - \mu\bar{y}^2 \\
\mu_{30} &= m_{30} - 3m_{20}\bar{x} + 2\mu\bar{x}^3 \\
\mu_{21} &= m_{21} - m_{20}\bar{y} - 2m_{11}\bar{x} + 2\mu\bar{x}^2\bar{y} \\
\mu_{12} &= m_{12} - m_{02}\bar{x} - 2m_{11}\bar{y} + 2\mu\bar{x}\bar{y}^2 \\
\mu_{03} &= m_{03} - 3m_{02}\bar{y} + 2\mu\bar{y}^3.
\end{aligned}
$$

- Normalize for scale-invariance: $\eta_{pq} = \dfrac{\mu_{pq}}{(\mu_{00})^{\gamma}}$

  - where $\gamma = (p+q)/2 + 1$ and $p + q \geq 2$

Hu-moments

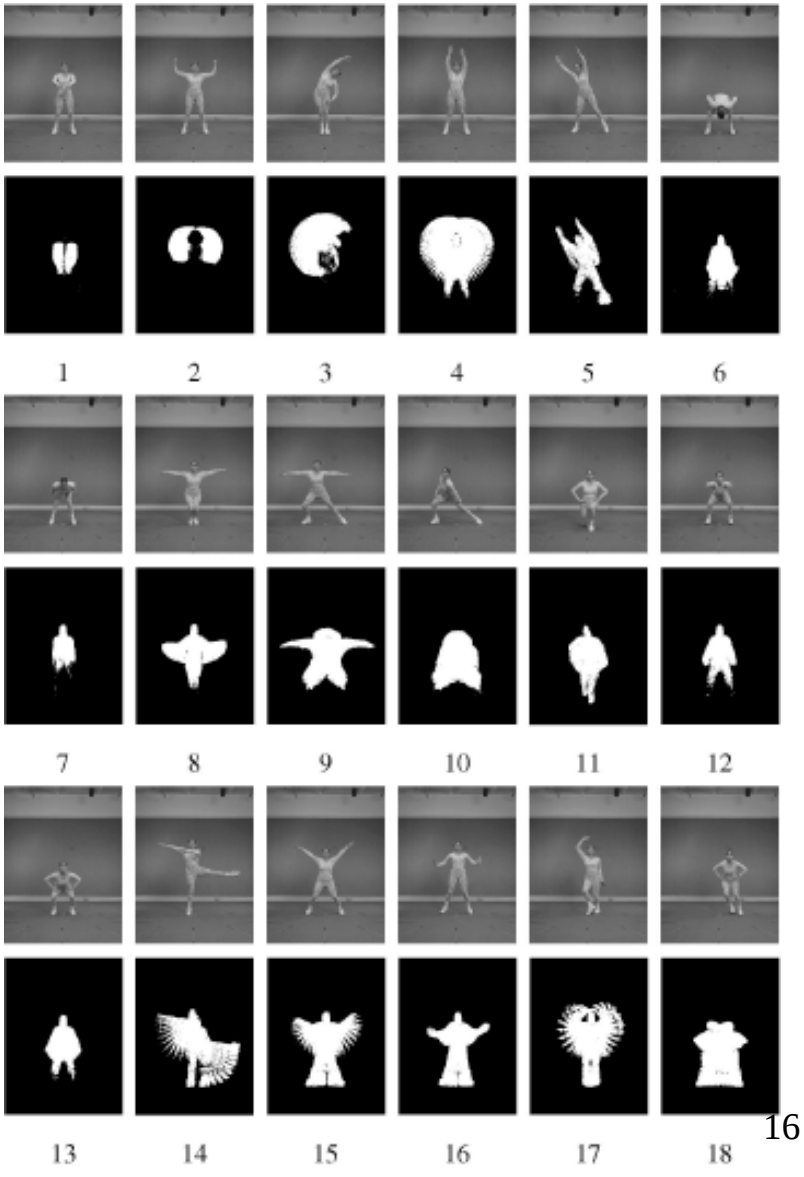
- The first seven orientation invariant Hu-Moments

$$
\begin{aligned}
\nu_1 &= \eta_{20} + \eta_{02} \\
\nu_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\nu_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\nu_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\nu_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \\
&\quad \cdot [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
\nu_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
&\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\nu_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned}
$$

- Hu-Moments are translation-, scale- and rotation-invariant.

Recognized Moves

- 18 aerobic exercises
- Several executions
- Seven views (-90 to +90 deg, 30deg increments)
- Results
  - With 1 camera: 12 out of 18 correct
  - With 2 cameras: 15 out of 18 correct



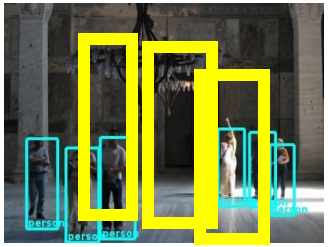
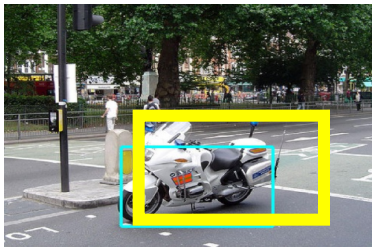
Example approaches

Boosting

Actions == space-time objects?

"stable-view" objects



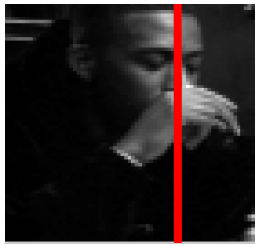
"atomic" actions

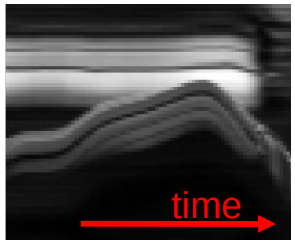


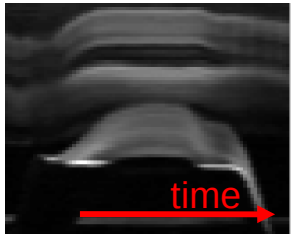


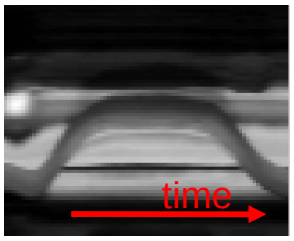
car exit | phoning | smoking | hand shaking | drinking

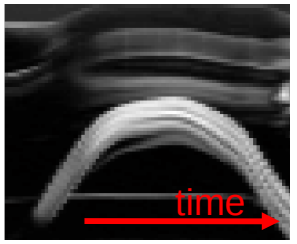
Take advantage of space-time shape





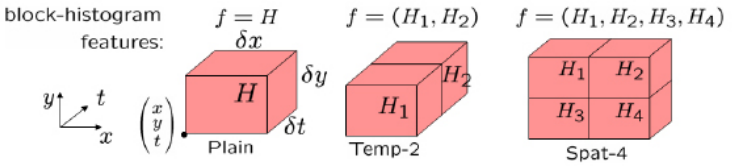


Temporal slice

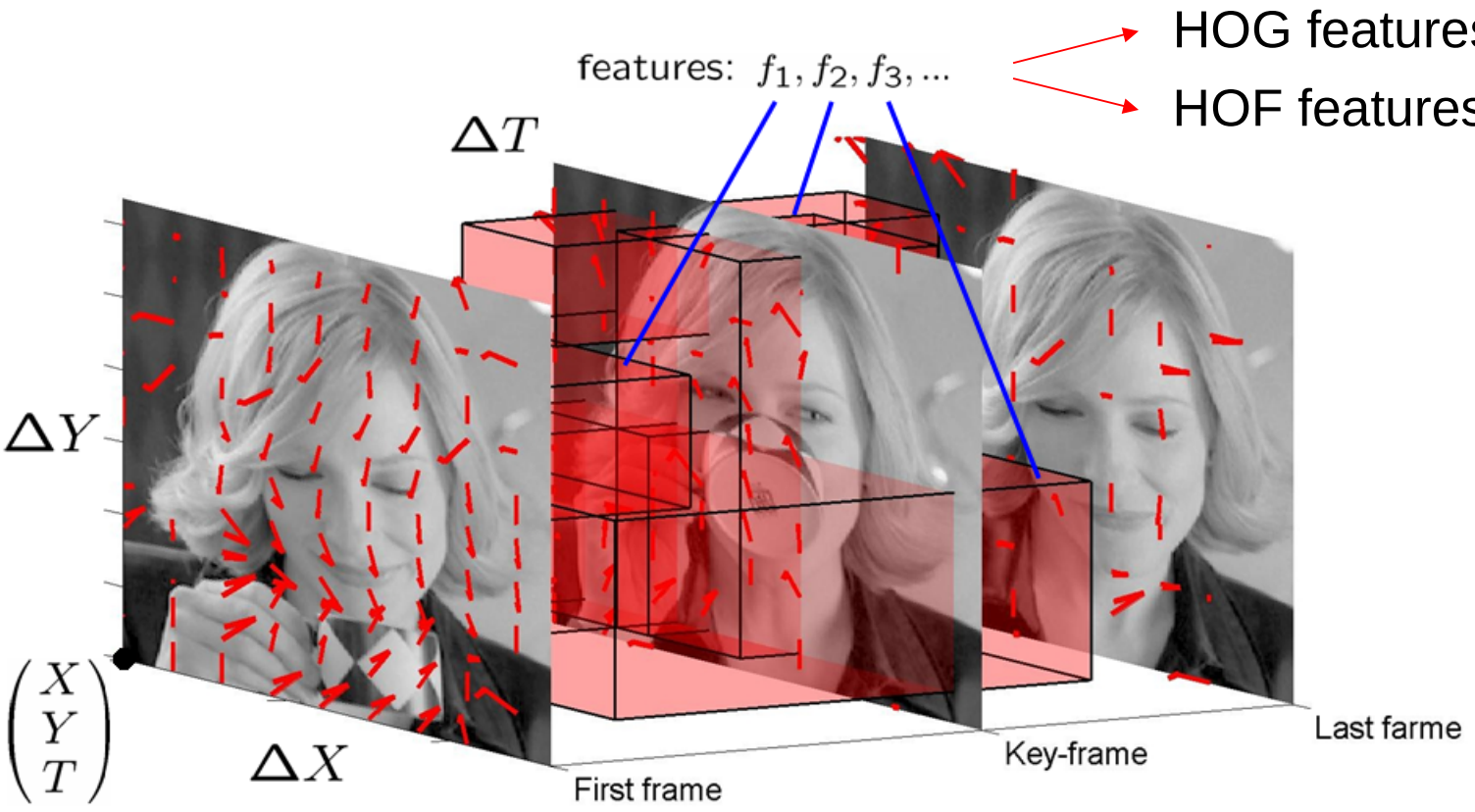
Action features [Laptev07]

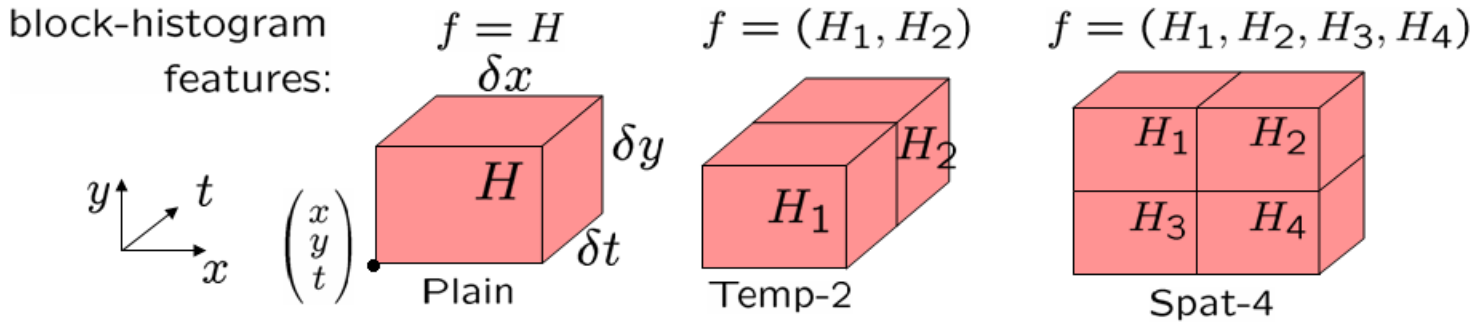
- Action volume = space-time cuboid region around the head (duration of action)

- Encoded with block-histogram features $f_\theta(.)$, $\theta = (x,y,t,dx,dy,dt,\beta,\varphi)$, defined by
- Location (x,y,t)
- Space-time extent (dx,dy,dt)
- Type of block (β)
  - β *el {Plain, Temp-2, Spat-4}*
- Type of histogram (φ)
  - Histogram of optical flow (HOF)
  - Histogram of oriented gradient (HOG)



Action features

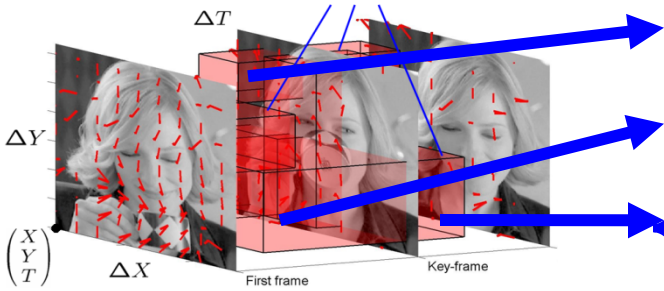


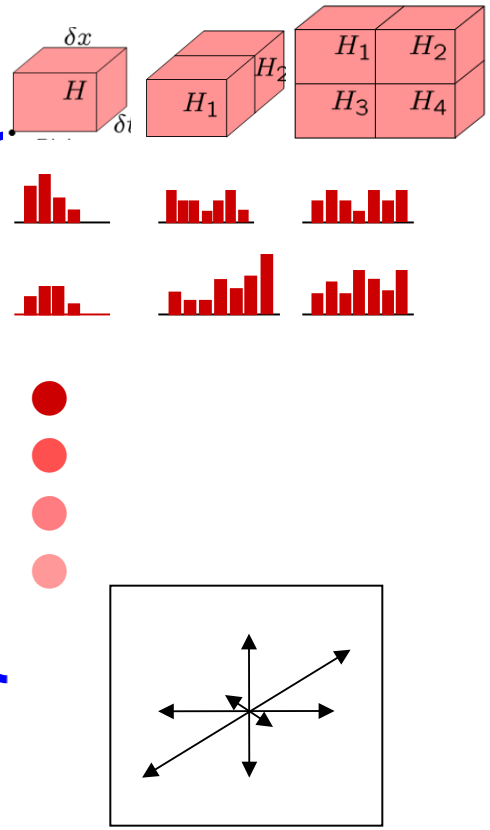


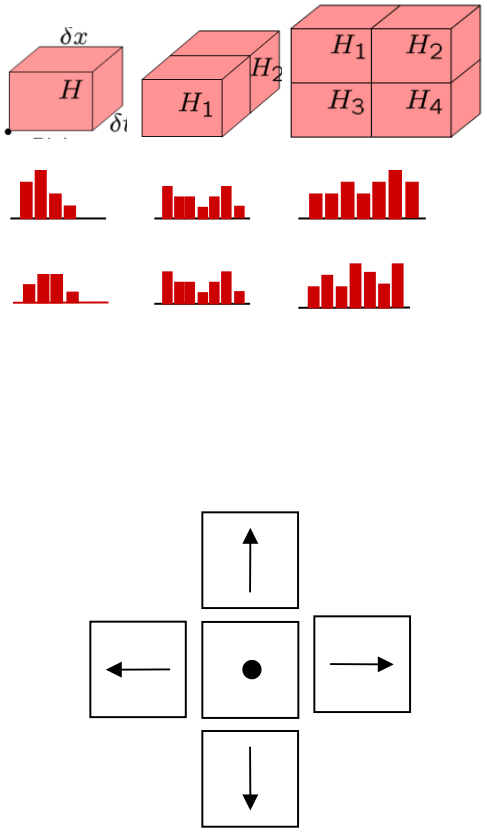
Histogram features

HOG: histograms of **oriented gradient**

HOF: histograms of **optic flow**







>10^6 possible features

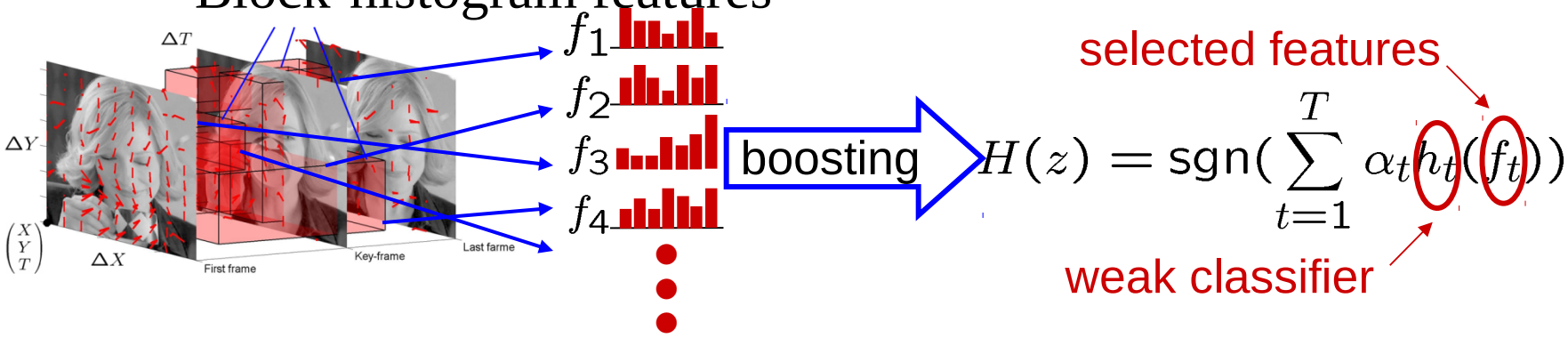
4 grad. orientation bins

4 OF direction bins
+ 1 bin for no motion

Action learning

- Use boosting method (eg. AdaBoost) to classify features within an action volume

- Features:
  - Block-histogram features



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

Action learning: Boosting

- A **weak classifier** *h* is a classifier with accuracy only slightly better than chance

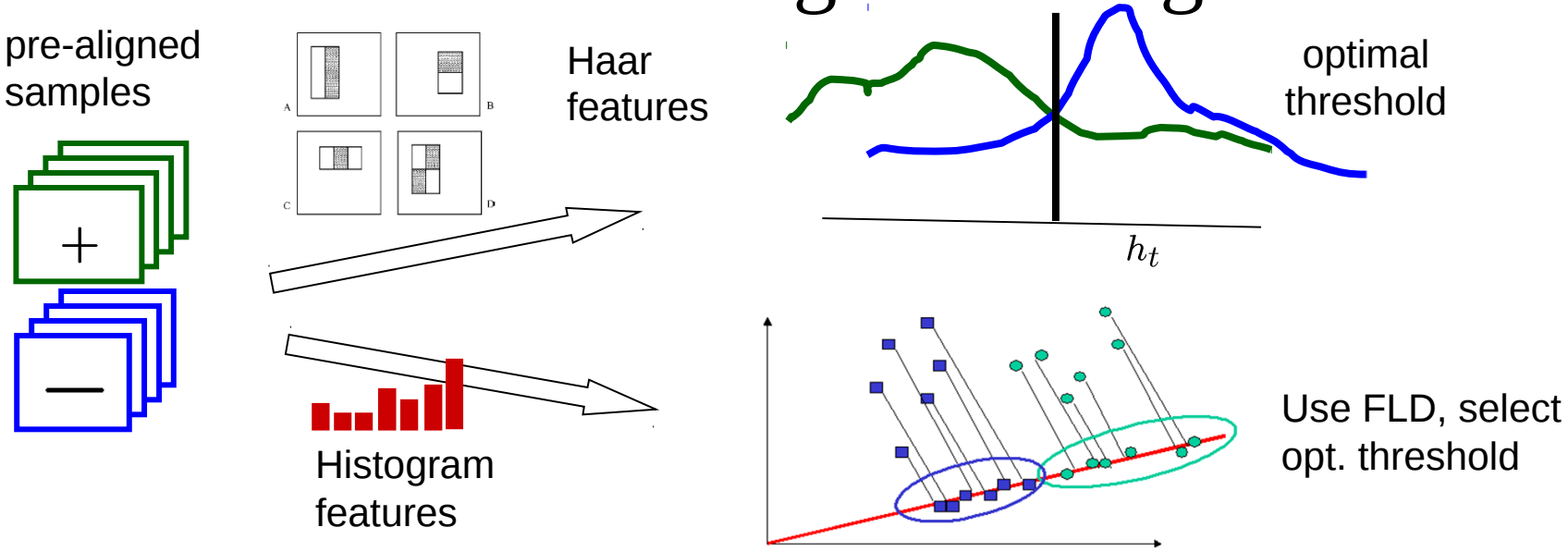
selected features

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

weak classifier

- Boosting: combine a number of weak classifiers so that the ensemble is arbitrarily accurate
  - Allows the use of simple (weak) classifiers without the loss if accuracy
  - Selects features and trains the classifier

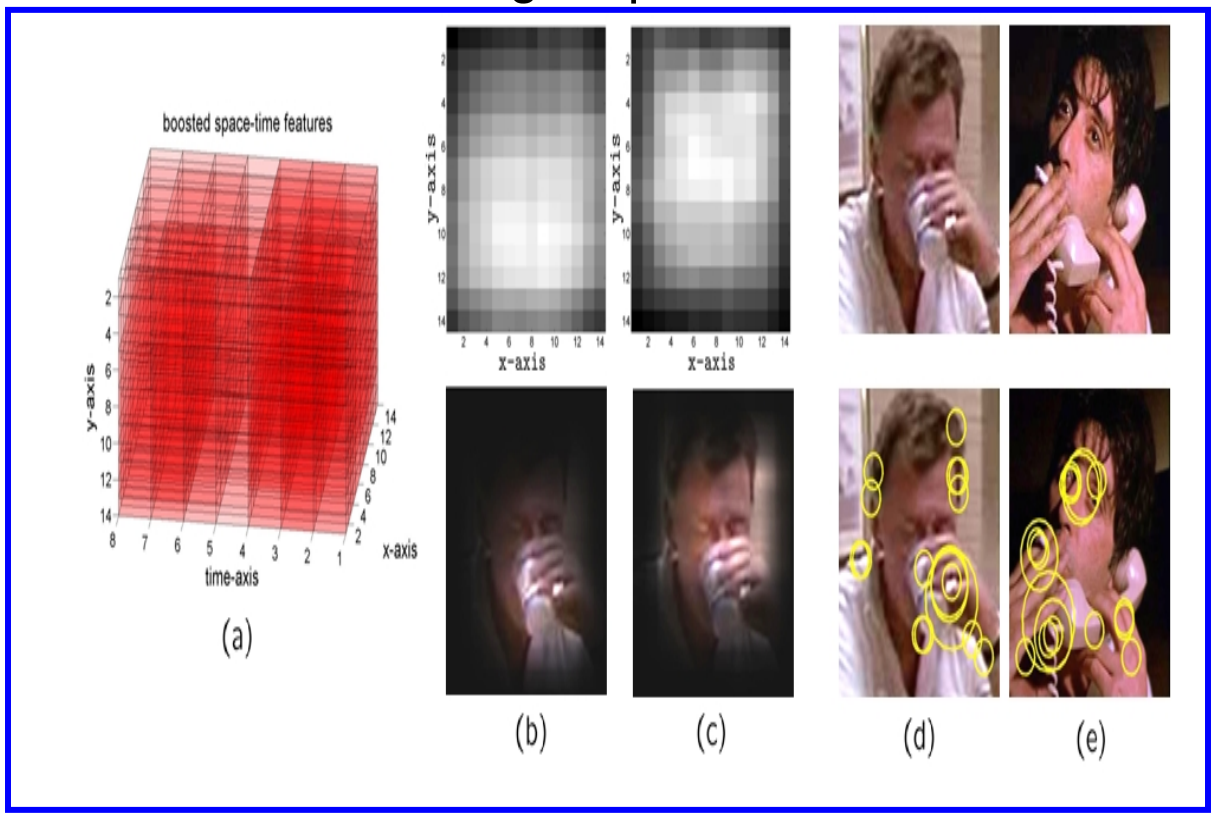
Action learning: Boosting

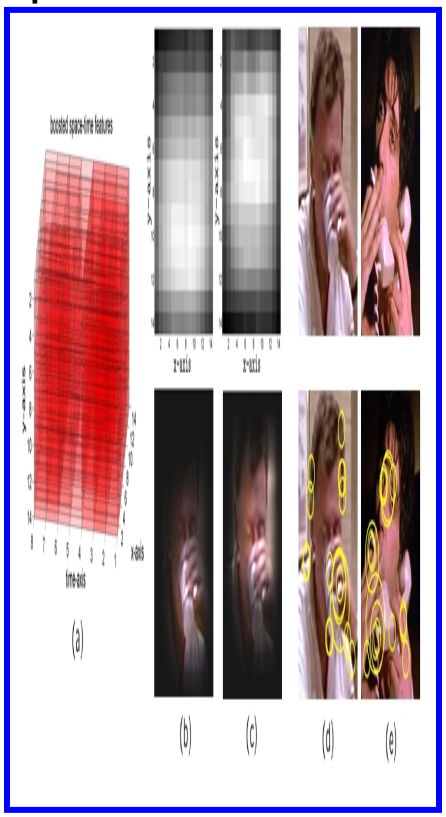


- Weak classifier h_t:
  - In case of one dimensional features
    - select an optimal decision threshold
    - E.g. for Haar-filter responses (Viola&Jones face detector)
  - Here: m-dimensional features
    - Project data on one dimension using Fisher's Linear Discriminant (FLD), then select optimal threshold in 1-D

Classifier properties

Training output: Accumulated feature maps





Space-time classifier (HOF)

Static keyframe classifier(HOG)

- Space-time classifier and static keyframe classifier might have complementary features

Keyframe priming

- Combination of static key-frame classifier with space-time classifier
  - Motivated by complementary properties of both classifiers
  - Bootstrap space-time classifier and apply it to keyframes detected by the keyframe detector (boosted space-time window classifier)

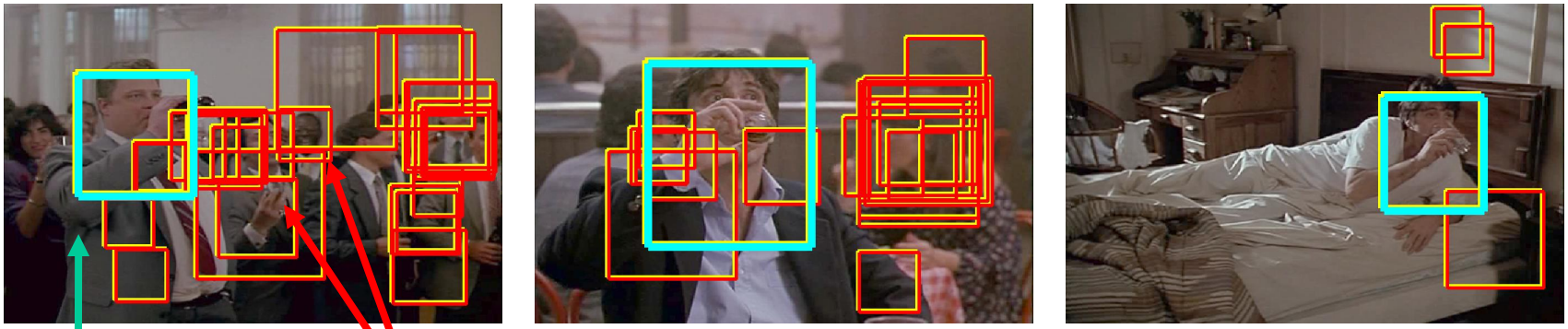
  - → Speeds up detection
  - → Combines complementary models

Keyframe priming

- Apply keyframe detector (HOG classification on single frame) to all positions, scales and frames while being set to a high false positive rate (10^{-3})
- Generate space-time blocks aligned with detected keyframes and with different temporal extent
- Run space-time classifier on each hypothesis

Keyframe priming

Training



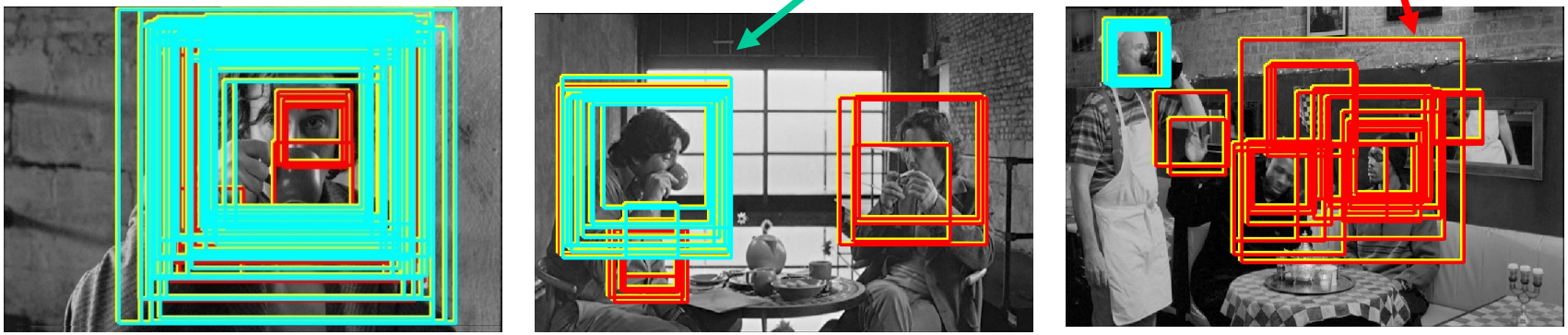
Positive training sample

Negative training samples

Keyframe-primed event detection

Keyframe detections

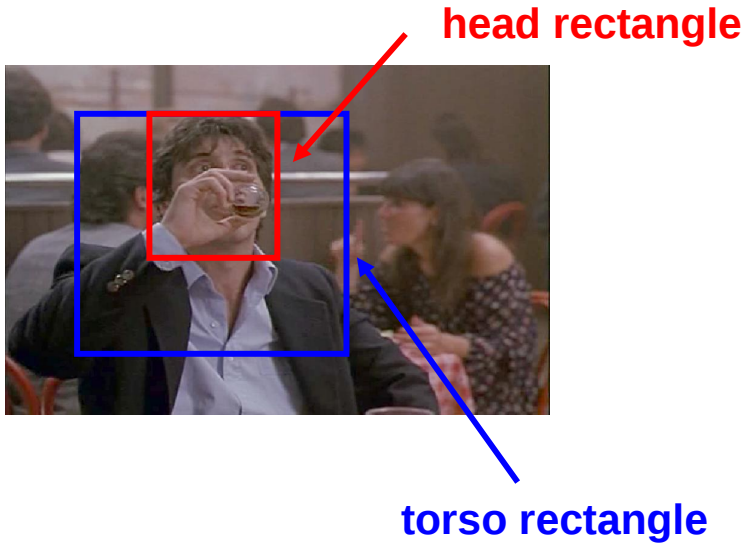
Test



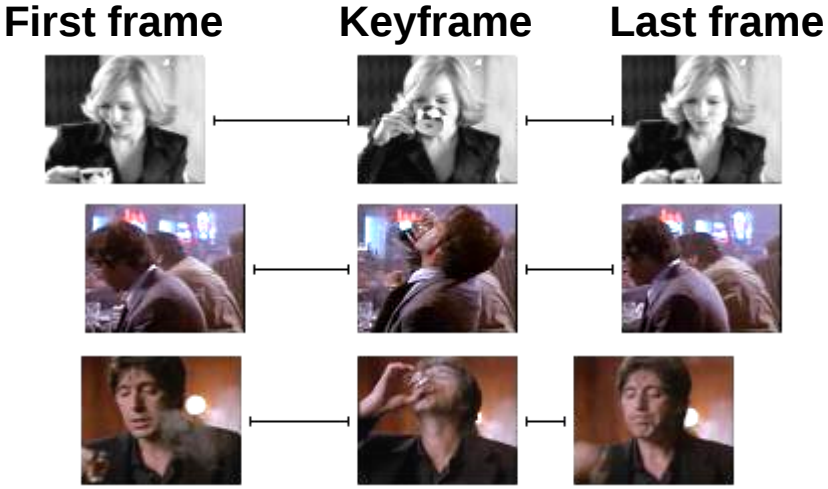
Action event dataset

- "Coffee and Cigarettes" dataset
  - 159 annotated "Drinking" samples
  - 149 annotated "Smoking" samples

Spatial annotation



Temporal annotation



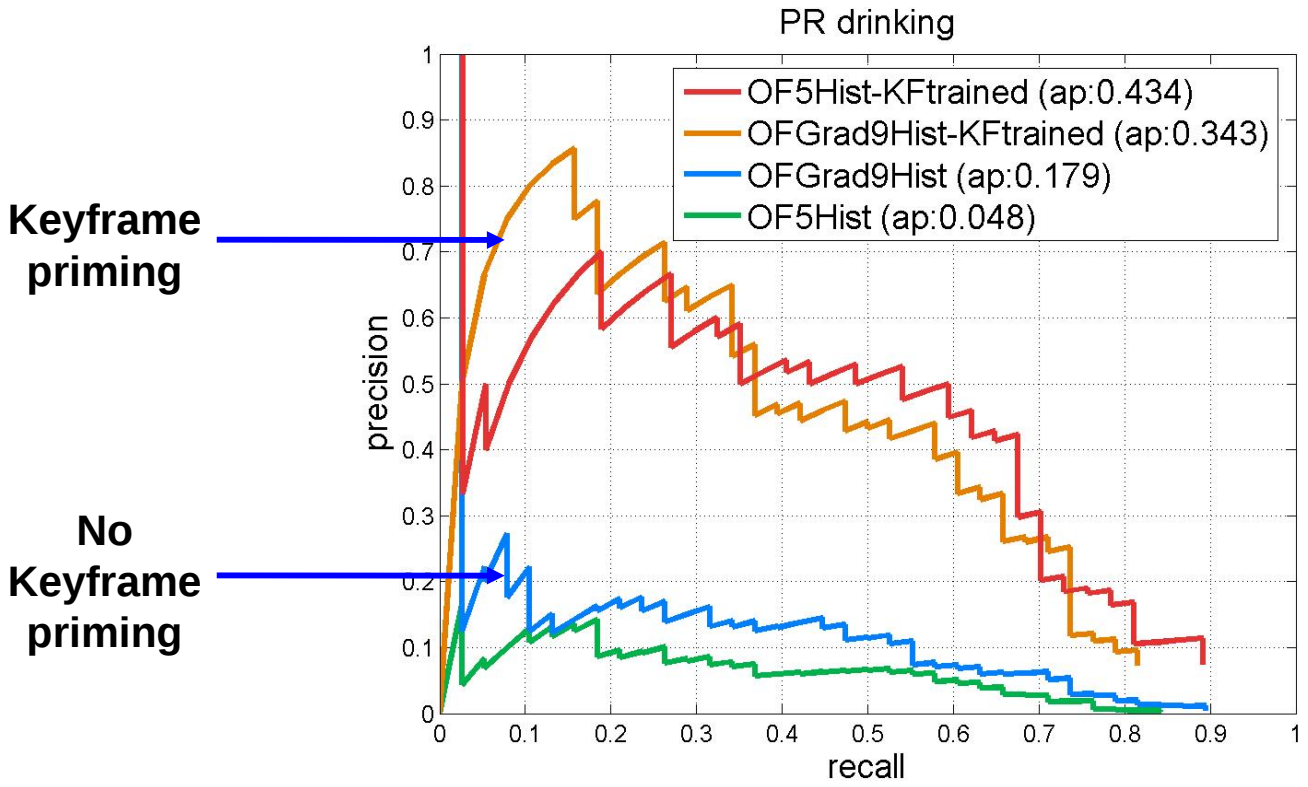
http://www.irisa.fr/vista/Equipe/People/Laptev/actiondetection.html

Action event dataset

- Challenges
  - High within-class variability of actions in terms of visual appearance
  - Similarity of both action classes in gross motion and posture of people

Action detection

- Test on 25min from “Coffee and Cigarettes” with 38 drinking actions
- No overlap with the training set in subjects or scenes
- → Keyframe priming is faster and leads to significant better results



Action detection

Example approaches

Bag-of-Words model

Action recognition in real-world videos

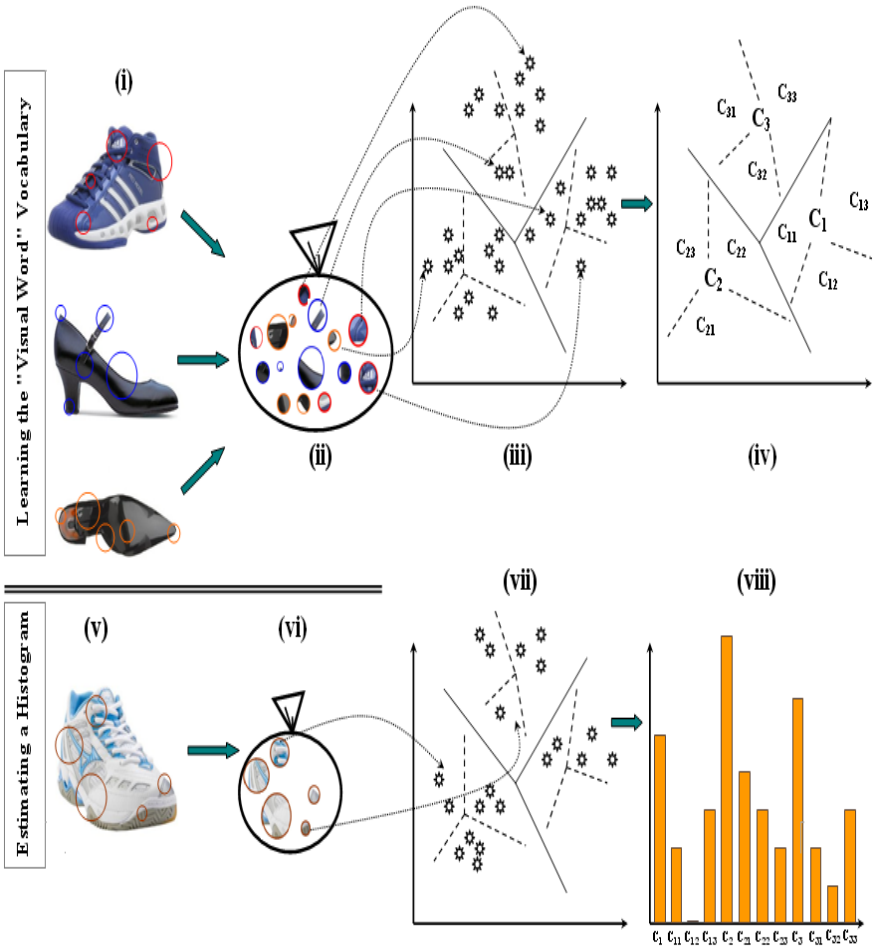
- So far
  - Few, simple action classes
- Robust detection and classification of all kinds of human actions needed [Laptev08]






Bag-of-Words (BoW) model

- „Visual Word“ vocabulary learning
  - Cluster local features
  - Visual Words = Cluster Means
- BoW feature calculation
  - Assign each local feature most similar visual word
  - BoW feature = Histogram of visual word occurances within a region

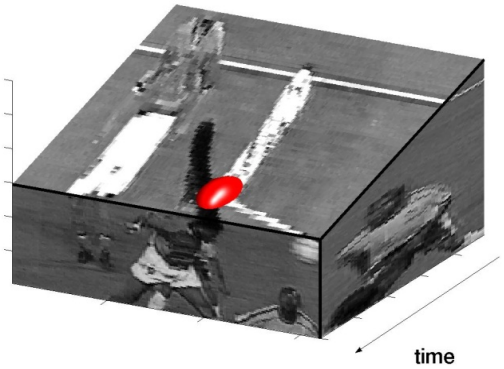


Space-Time Features: Detector

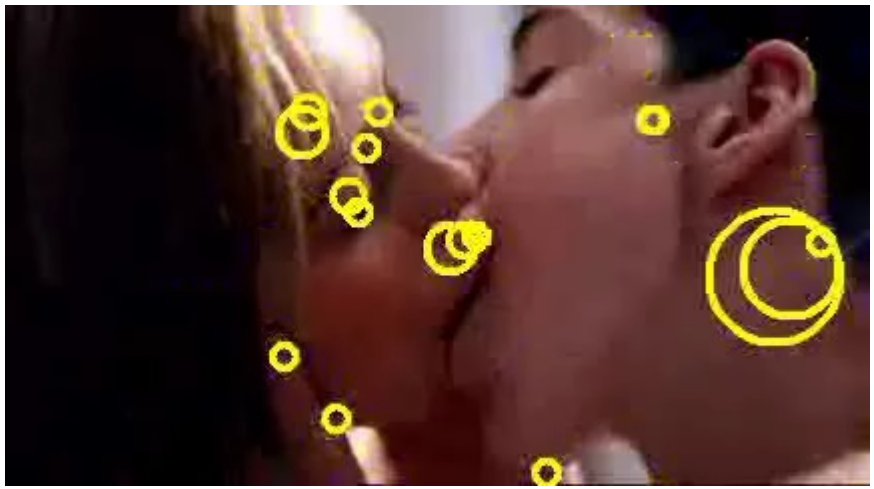
[Laptev, IJCV 2005]

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

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

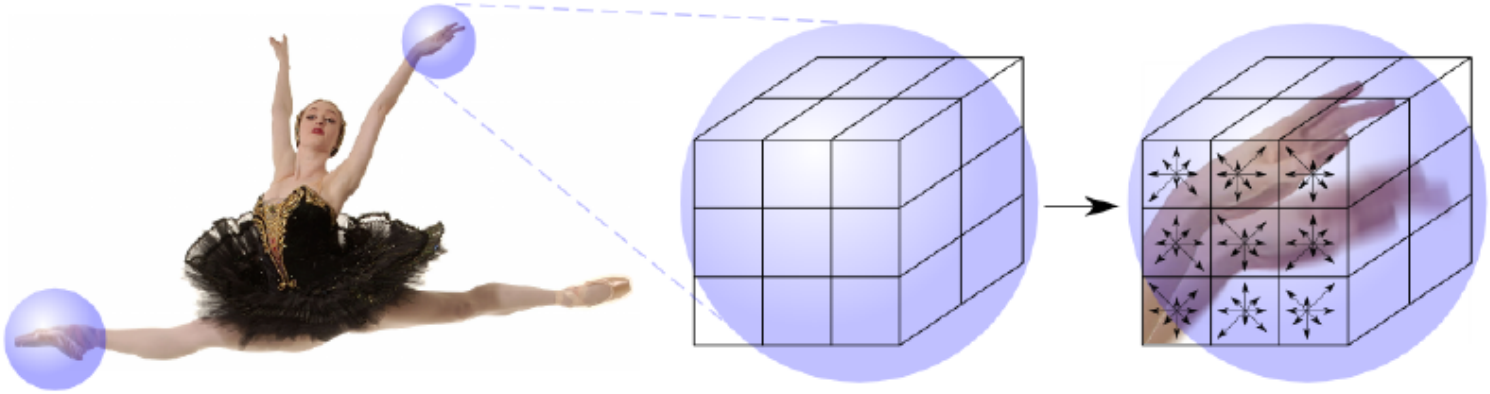


$$(\sigma^2,\tau^2) = \mathcal{S}\times\mathcal{T},\ \mathcal{S} = 2^{\{2,\ldots,6\}}, \mathcal{T} = 2^{\{1,2\}}$$

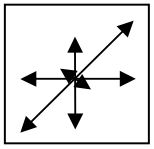


Space-Time Features: Descriptor

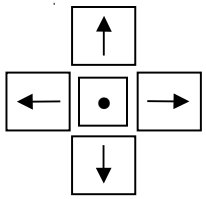
Multi-scale space-time patches
from corner detector



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

Space-Time Features: Descriptor

- Compute histogram descriptors of space-time volumes in neighborhood of detected points:
  - Compute a 4-bin HOG for each cube in 3x3x2 space-time grid
  - Compute a 5-bin HOF for each cube in 3x3x2 space-time grid
- Size of each volume related to detection scales

$$\Delta_x = \Delta_y = 2k\sigma \qquad \Delta_t = 2k\tau$$

Action classification

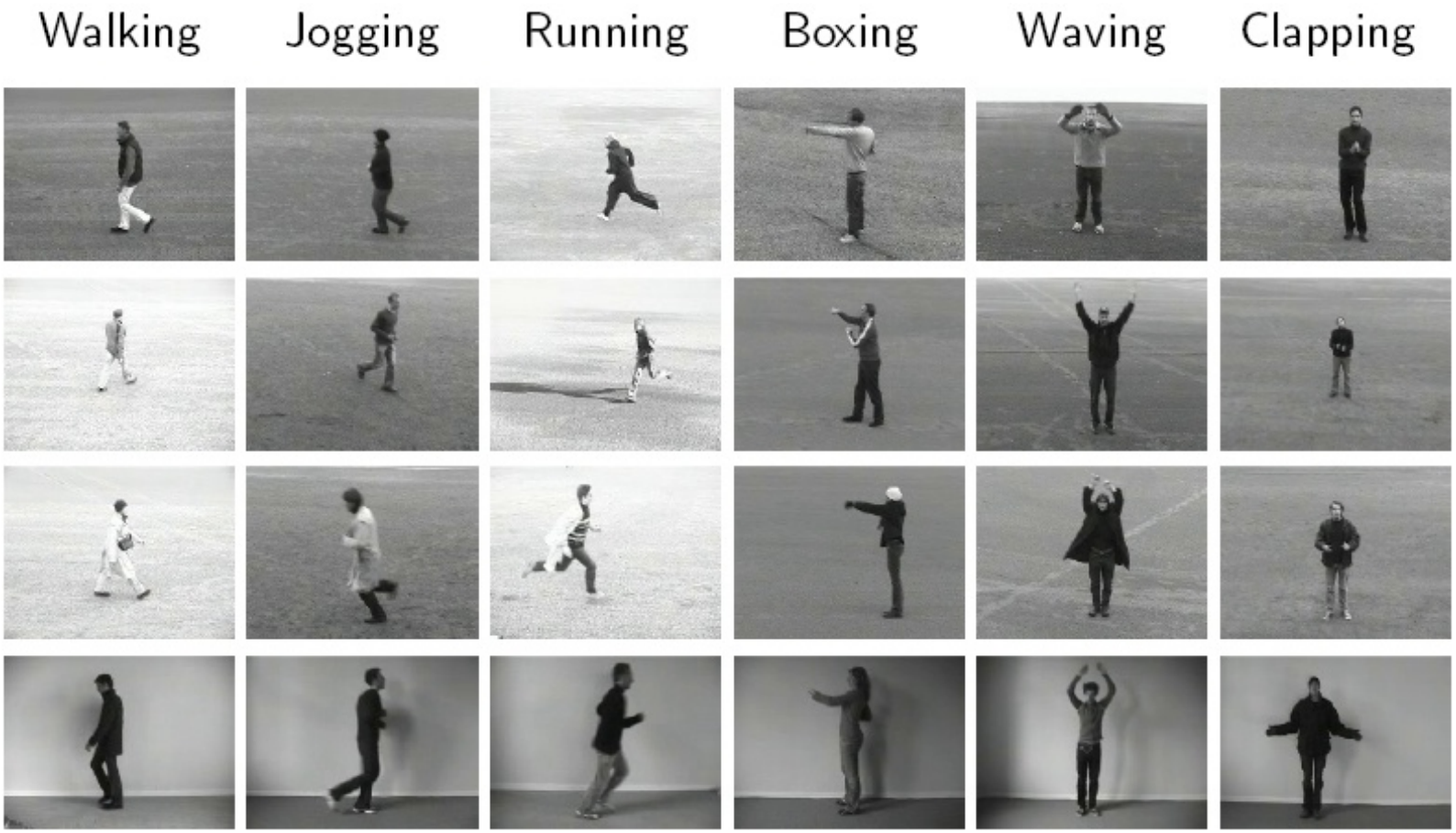
- Spatio-temporal Bag-of-Words (BoW)

  - Build Visual vocabulary of local feature representations using k-means clustering
  - Assign each feature in a video to nearest vocabulary word
  - Compute histogram of visual word occurrences over space time volume of a video sequence

- SVM classification
  - Combine different feature types using multichannel χ^2 Kernel
  - One-against-all approach in case of multi-class classification

Results on KTH actions dataset

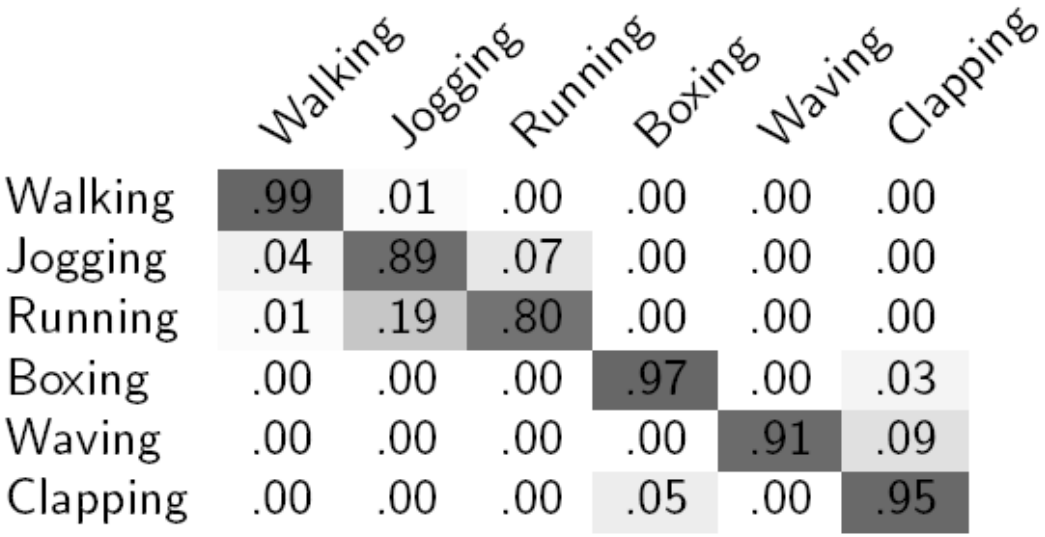
- Examples of all six classes and all four scenarios



Results on KTH actions dataset

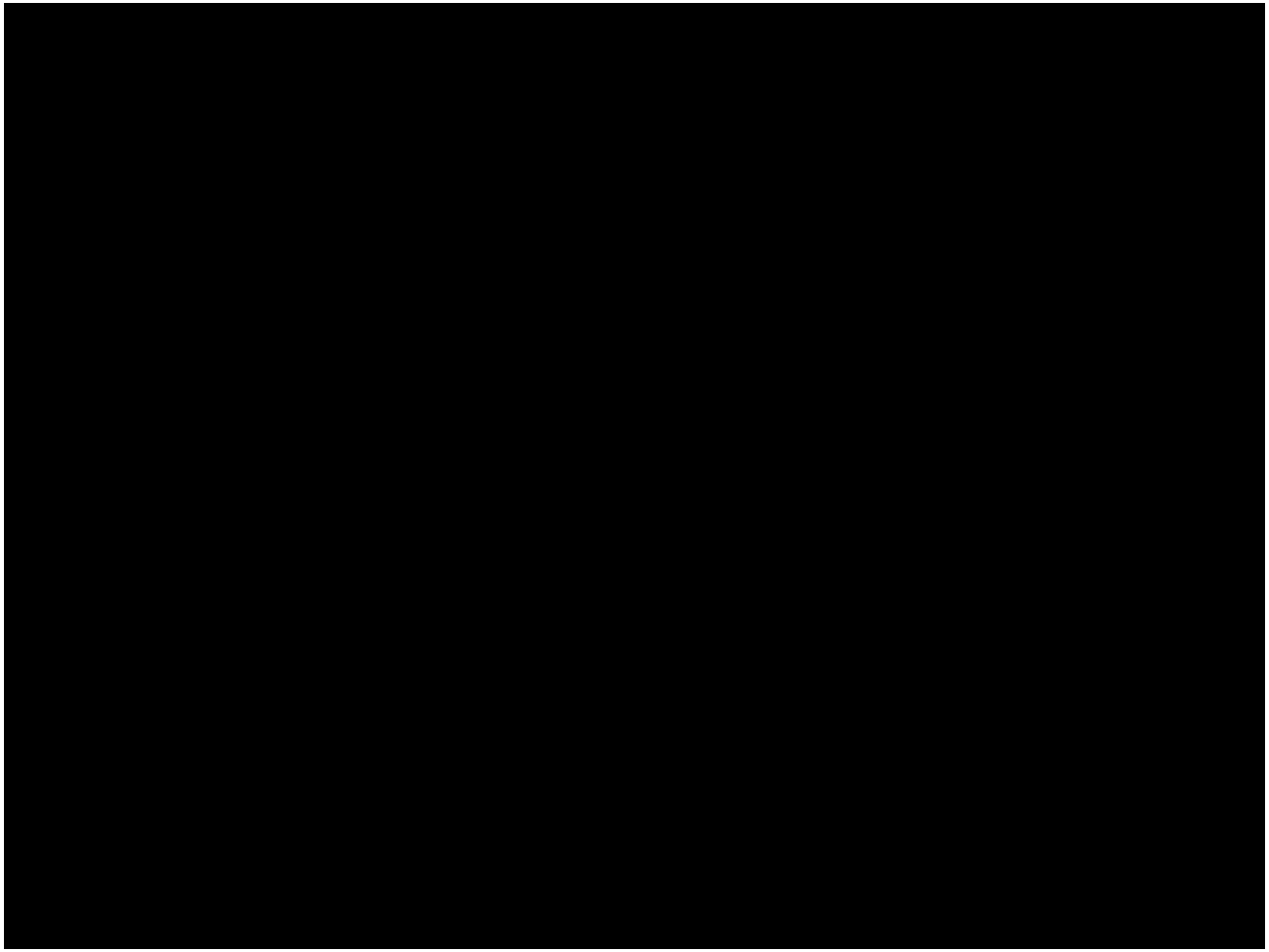
Method	Schuldt et al.	Niebles et al.	Wong et al.	Nowozin et al.	ours
Accuracy	71.7%	81.5%	86.7%	87.0%	**91.8%**

Average class accuracy



Confusion matrix

Results on HOHA dataset



Summary

- Techniques from objects detection/recognition work very well for action recognition
- Actions can be looked at as “space-time objects”

Approaches:

- Temporal Templates to classify aerobic movements
  - MEI, MHI, scale-, translation- and rotation-invariant features

- Boosting with motion and appearance features to detect and classify smoking and drinking actions

- Bag-of-Words model for action recognition in movies

References

F. Bobick, J. Davis. The Recognition of Human Movement Using Temporal Templates. IEEE PAMI, Vol. 23, No. 3, March 2001

I. Laptev & P. Pérez. Retrieving actions in movies. IEEE International Conference on Computer Vision, 2007

P. Viola & M. Jones. Robust Real-Time Face Detection. International Journal of Computer Vision, 2004

R. Schapire, Y. Freund, P. Bartlett & WS. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. ICML, 1997

I. Laptev, M. Marszalek, C. Schmid & B. Rozenfeld. Learning realistic human actions from movies. CVPR, 2008

I. Laptev. On Space-Time Interest Points. International Journal of Computer Vision, 2005