AARHUS
UNIVERSITET

# Computer Vision & Machine Learning

Alexandros Iosifidis
@
Department of Electrical and Computer Engineering
Aarhus University

# Human action recognition

What is a human action?



KTH dataset

# Human action recognition

What is a human action?

Weizmann dataset

# Human action recognition

What is a human action?








Weizmann dataset

# Human action recognition

What is a human action?

Weizmann dataset

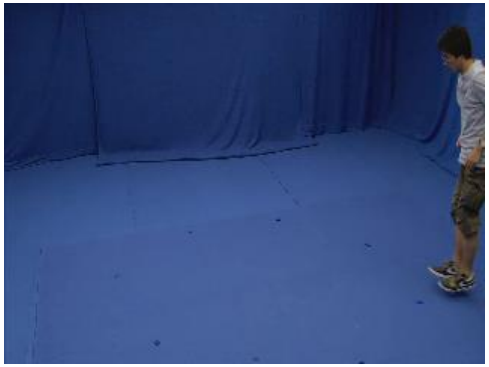
# Human action recognition

What is a human action?

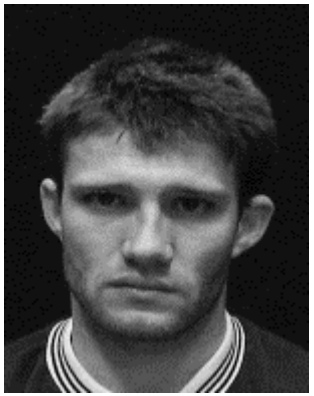
Hollywood movie dataset

# Human action recognition

What is a human action?







Other examples

# Human action recognition

Issues:

- Intra-class variability (background and styles)









Other examples

# Human action recognition

Issues:

- Intra-class variability (viewing angles)




Other examples

# Human action recognition

Issues:

- Position and proportion in the visual data (frames)

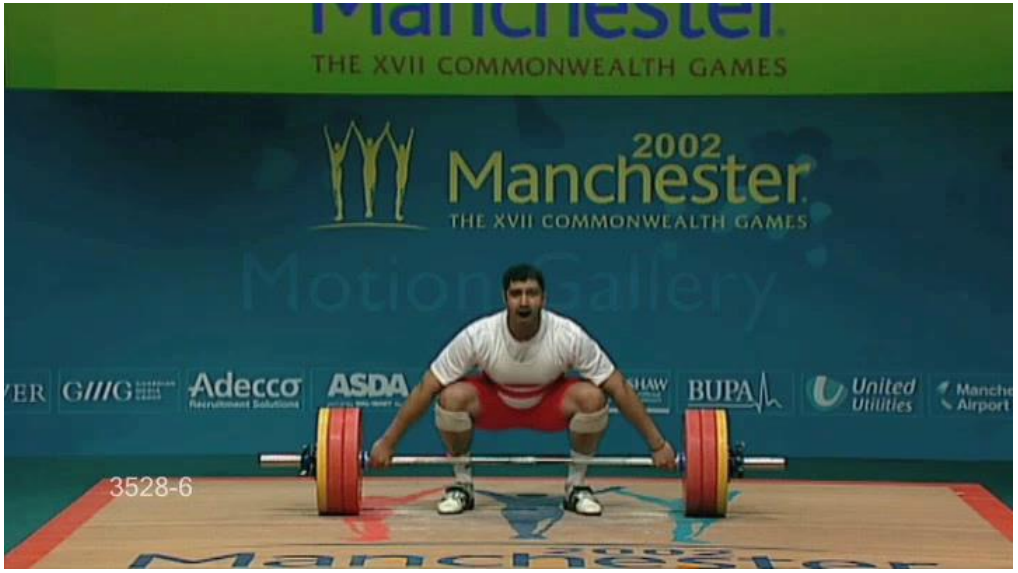



Other examples

# Human action recognition

Issues:

- Action duration


Other examples

# Human action recognition

Problems researched:

- Single-view action recognition (1990-2006): one camera, simple background, simple actions, (usually) one person actions
- Multi-view human action recognition (2000-present): multiple cameras, simple background, simple actions (usually) one or two person actions
- View-invariant human action recognition (2000-present): one camera, simple background, simple actions
- Action recognition in the wild (2006-present): (usually) one camera, cluttered background, one or many person actions
- Cross-view action recognition (2000-present): learning on one viewing angle and testing on another

# Human action recognition

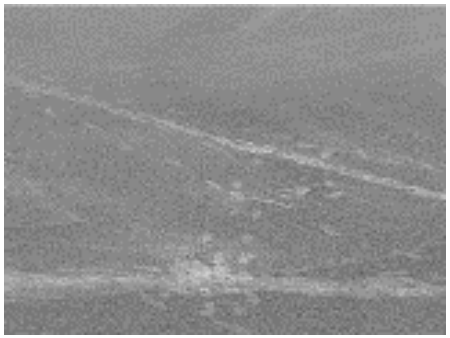
(Rough) categorization of methods:

- Methods exploiting global (body) information: Human body silhouettes
- Methods exploiting local video information: (Space-Time) Interest points and local image descriptors
- Methods based on neural networks: Raw video data analysis

# HAR based on global body information

Simple background, simple actions, (usually) one person actions



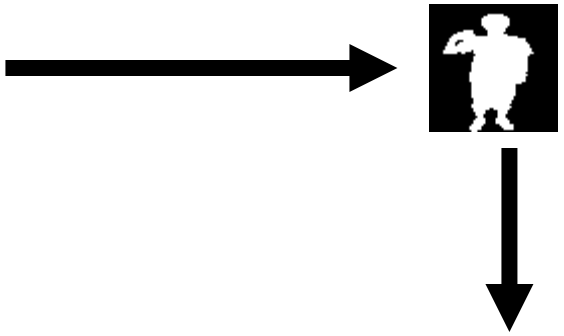




# HAR based on global body information

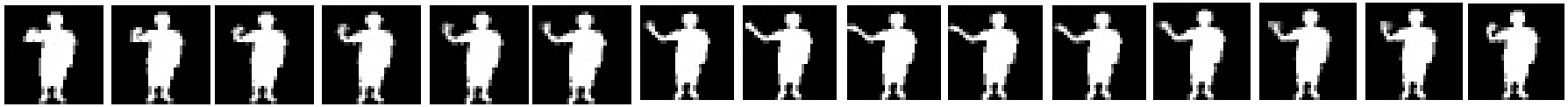
Simple background, simple actions, (usually) one person actions

Make easy the pre-processing steps!

Background subtraction, cropping and rescaling



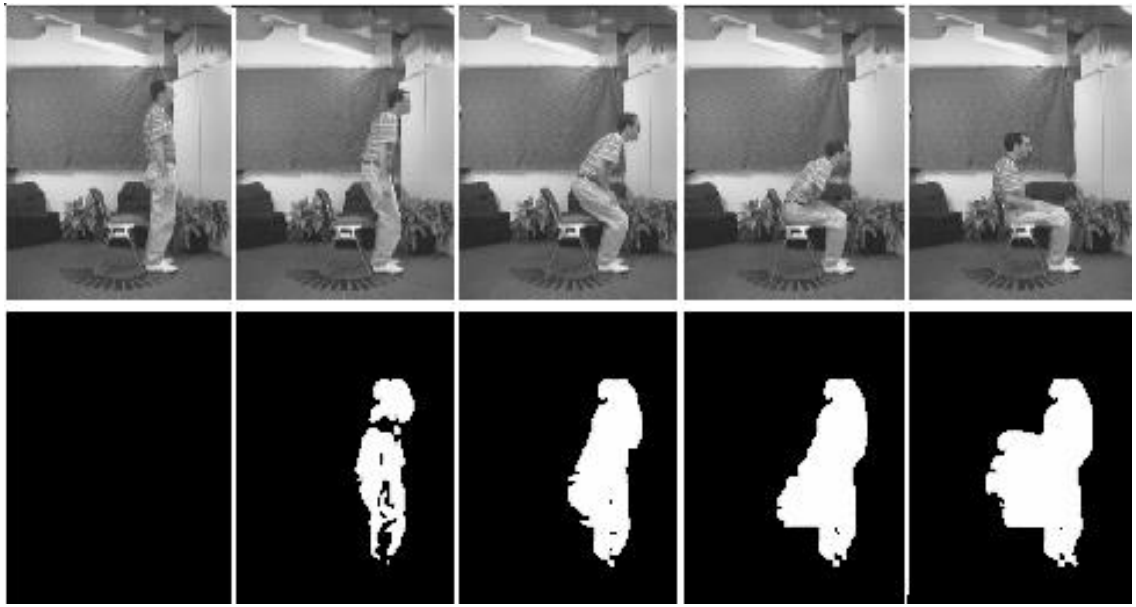
Action description: a set of consecutive human body poses



# HAR based on global body information

How to combine the various human body poses?

- Motion Energy Image (MEI)



# HAR based on global body information

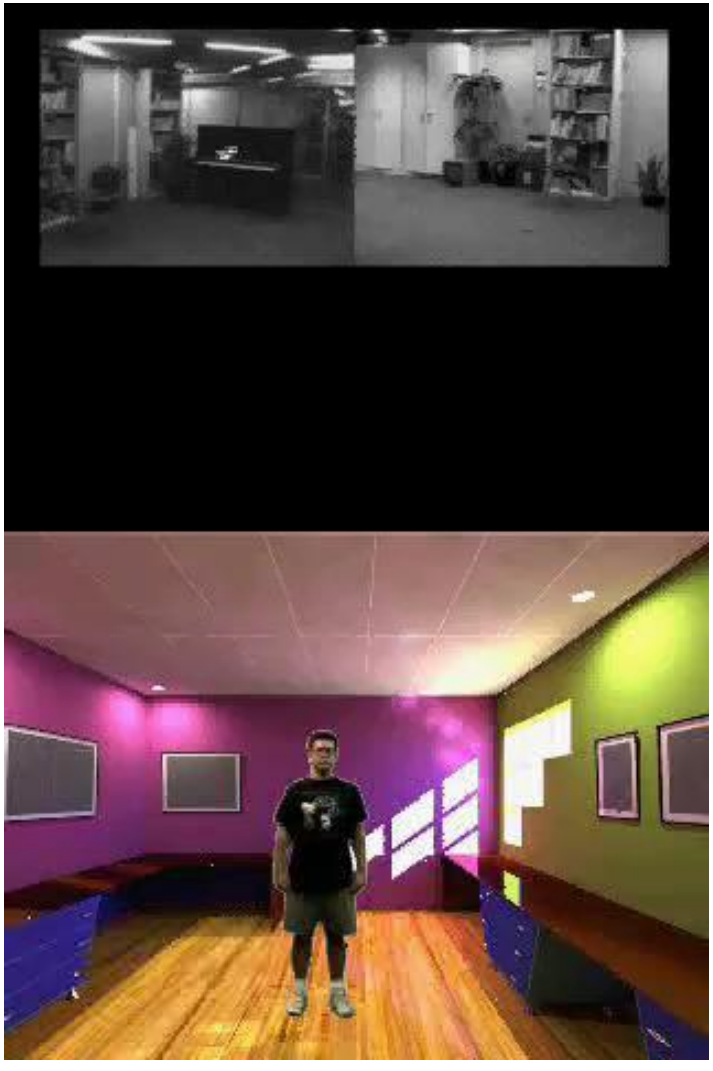
How to combine the various human body poses?

- Motion History Image (MHI)

# HAR based on global body information

How to combine the various human body poses?

- Motion History Image (MHI)

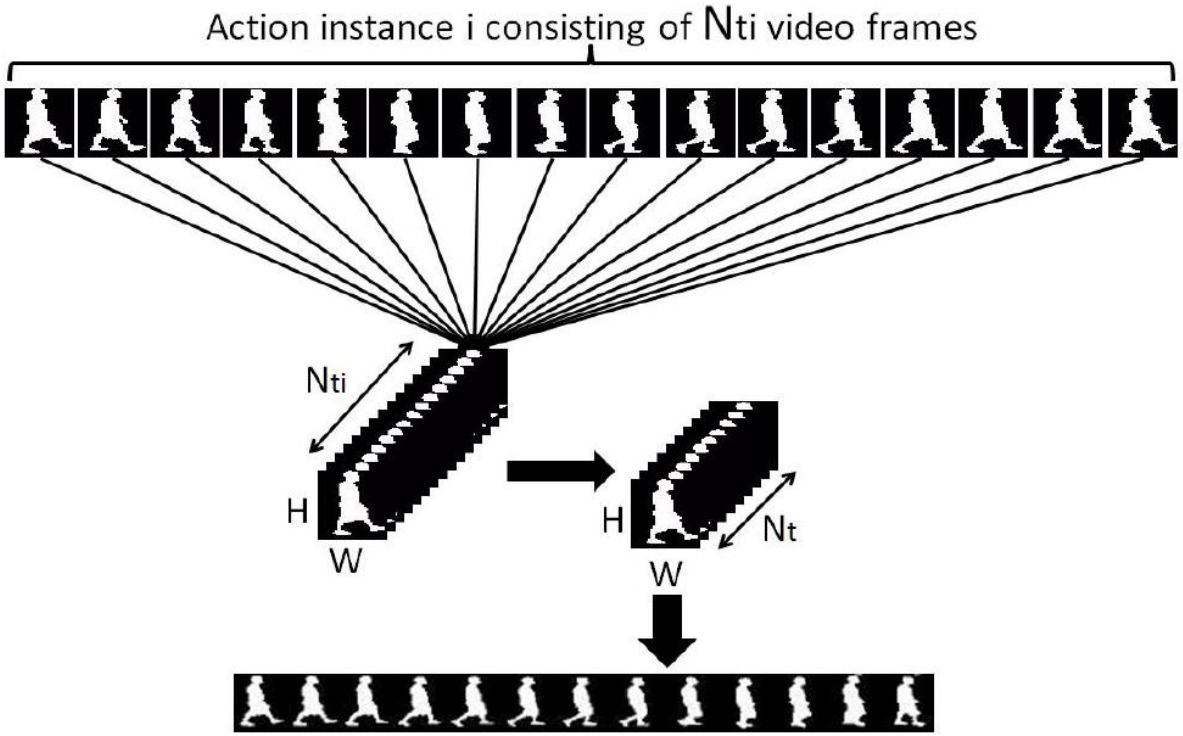


Demo

# HAR based on global body information

How to combine the various human body poses?

- Action Image (it requires scaling in the time domain too)



# HAR based on global body information

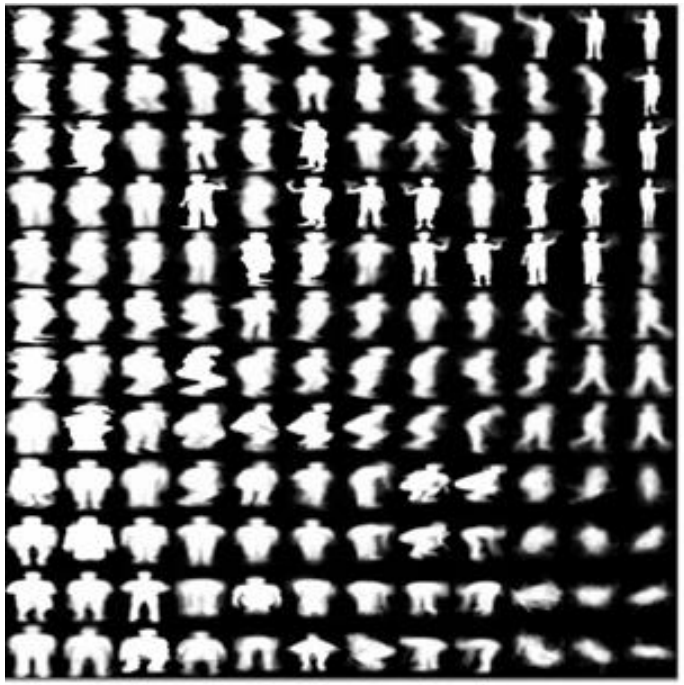
How to combine the various human body poses?

- Space-Time Volume

# HAR based on global body information

How to combine the various human body poses?

- Define a set of prototype poses by clustering the human body silhouettes of the training videos. These prototypes correspond to “dynamic poses”, and are called dynemes.



# HAR based on global body information

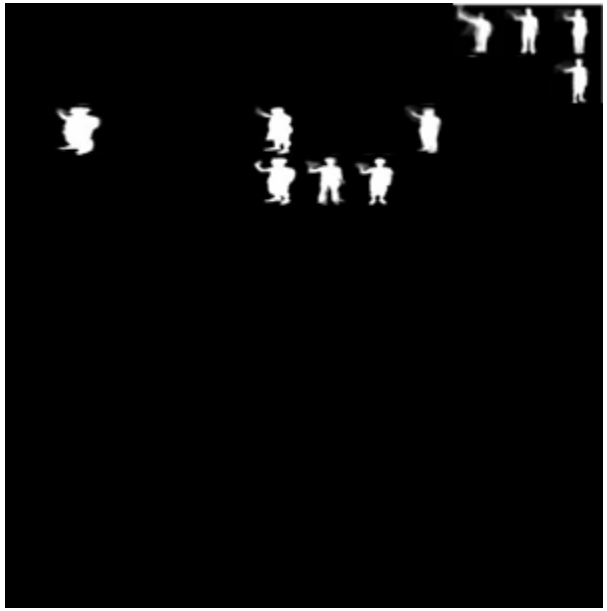
How to combine the various human body poses?

- Define a set of prototype poses by clustering the human body silhouettes of the training videos. These prototypes correspond to “dynamic poses”, and are called dynemes.
- A new video is described by the corresponding human body poses

# HAR based on global body information

How to combine the various human body poses?

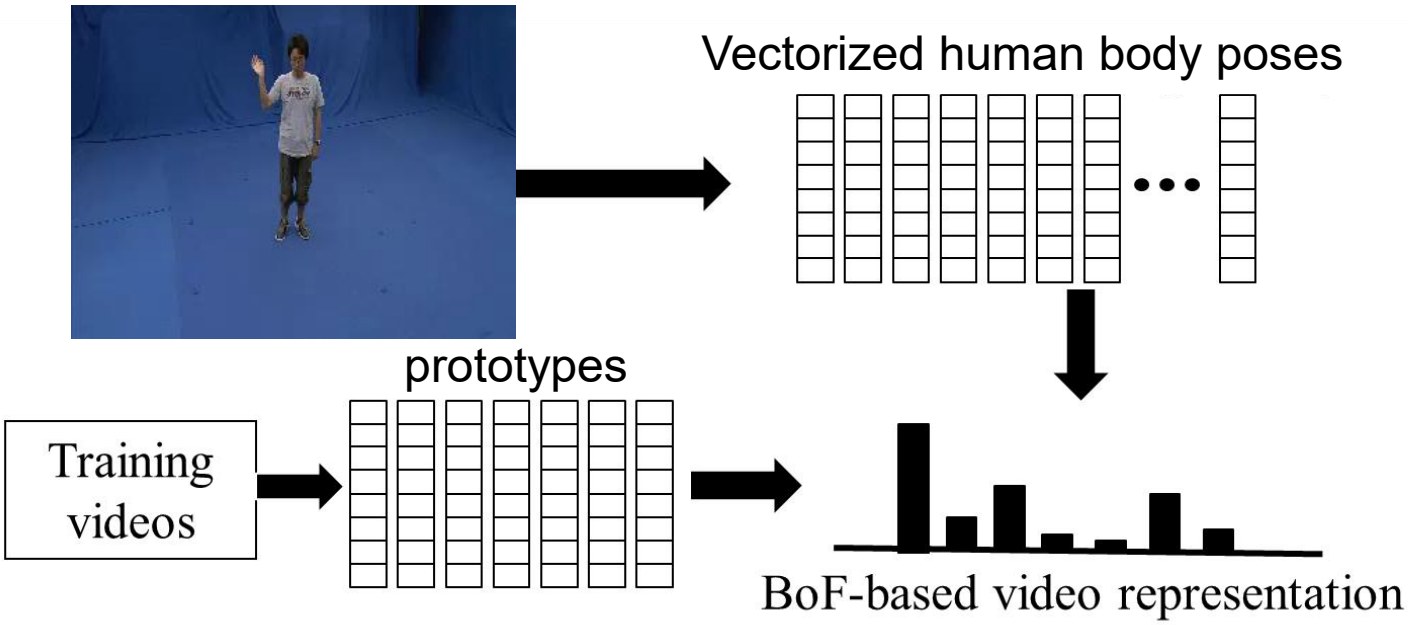
- Define a set of prototype poses by clustering the human body silhouettes of the training videos. These prototypes correspond to “dynamic poses”, and are called dynemes.
- A new video is described by the corresponding human body poses
- Assign (with soft or hard assignment) the body poses to dynemes → Histogram of human body poses (using hard or soft quantization)



# HAR based on global body information

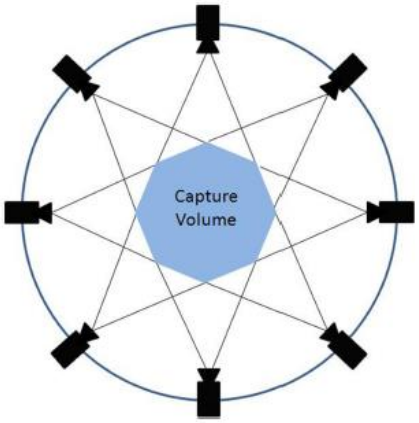
How to combine the various human body poses?

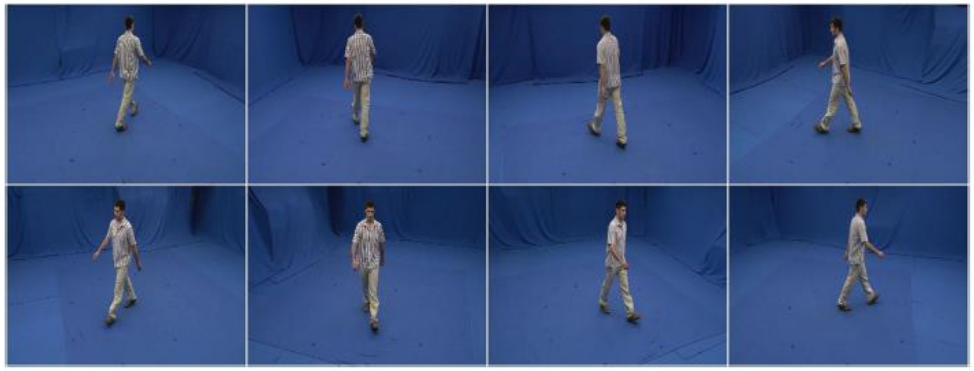
- Dyneme-based action recognition

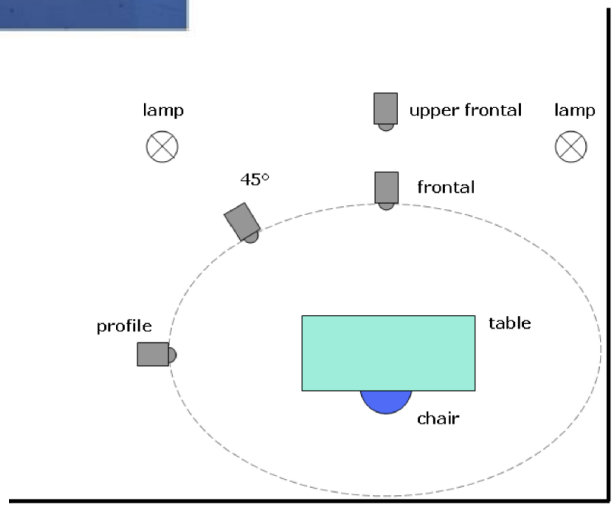


# HAR based on global body information

In case where multiple cameras are available:





# HAR based on global body information

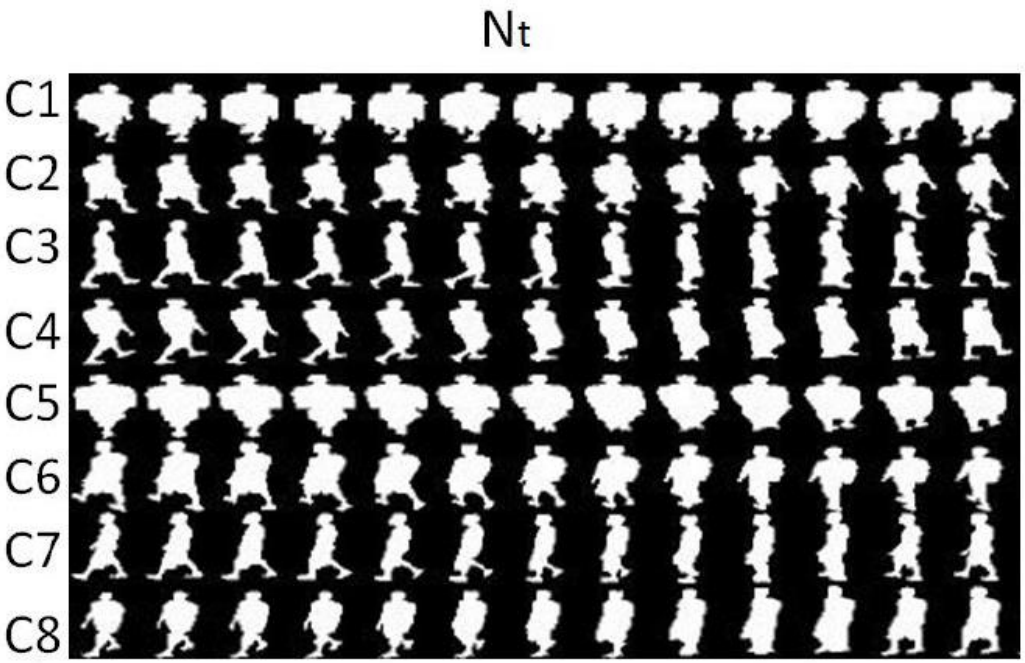
In case where multiple cameras are available:

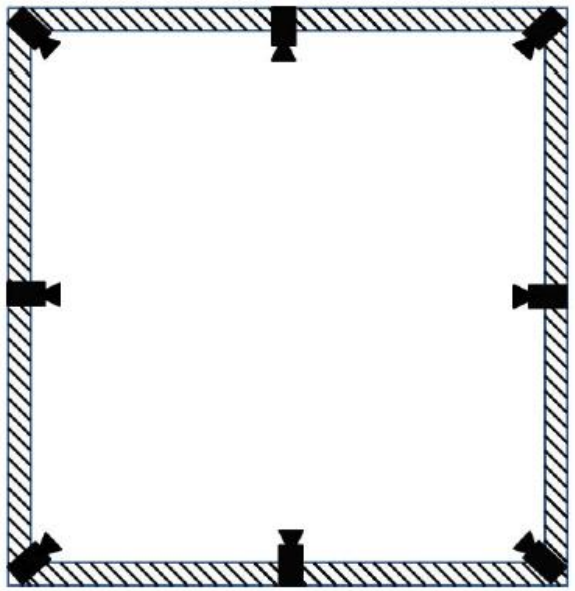
- Multi-view human body pose

# HAR based on global body information

In case where multiple cameras are available:

- Multi-view Action Image

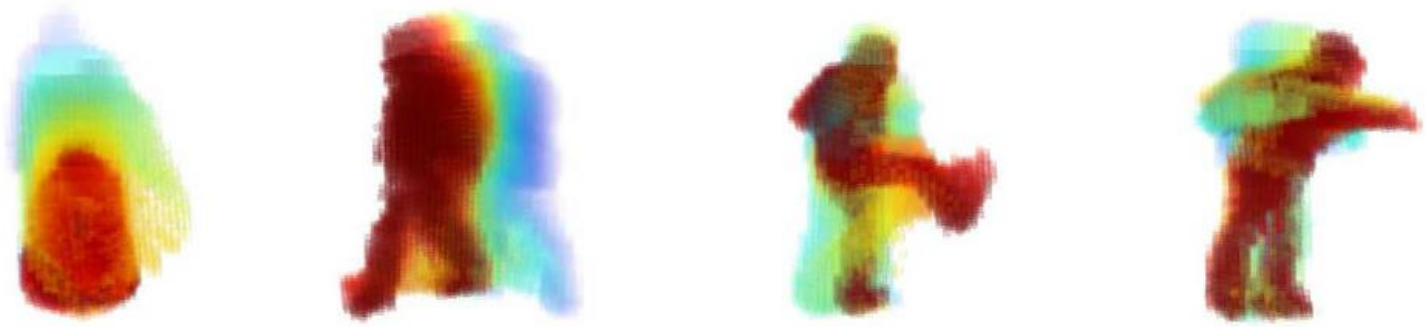


# HAR based on global body information

In case where multiple cameras are available:

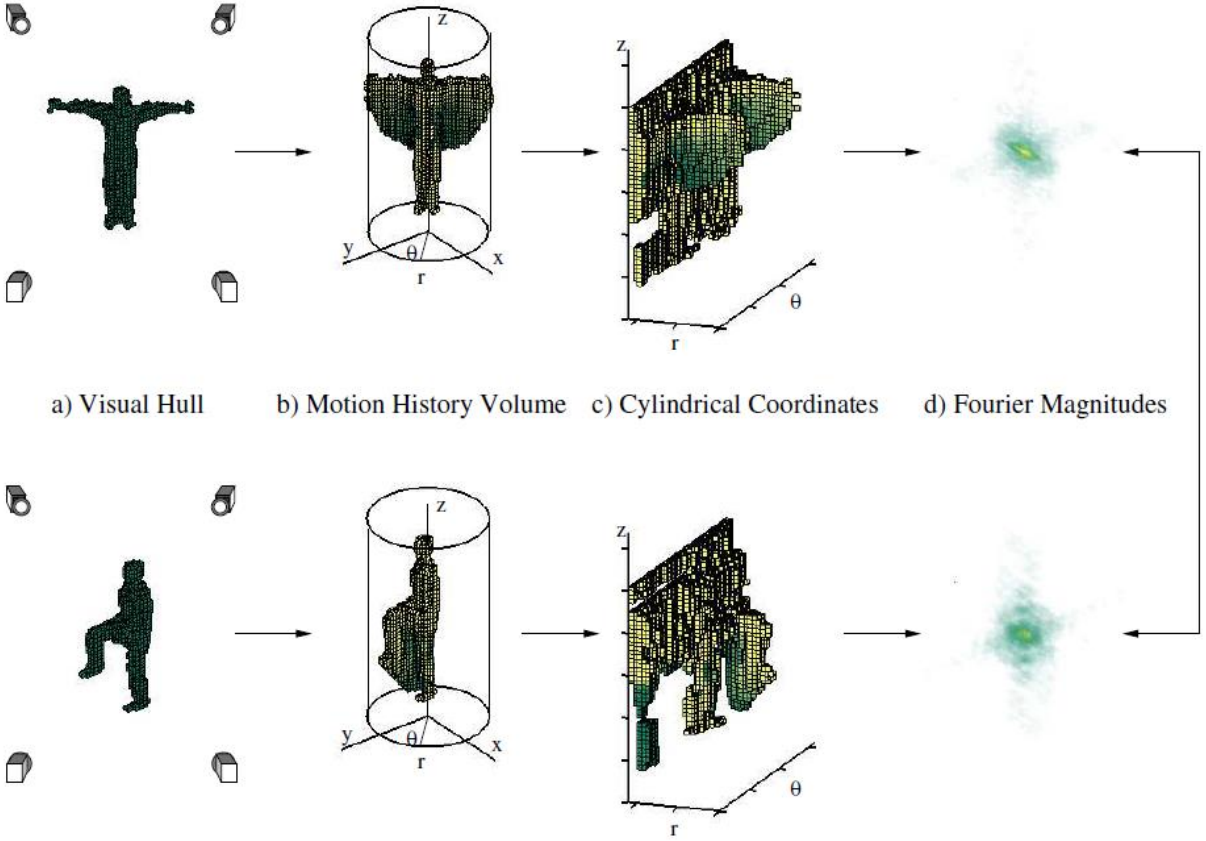
- Motion History Volume



# HAR based on global body information

In case where multiple cameras are available:

- Motion History Volume

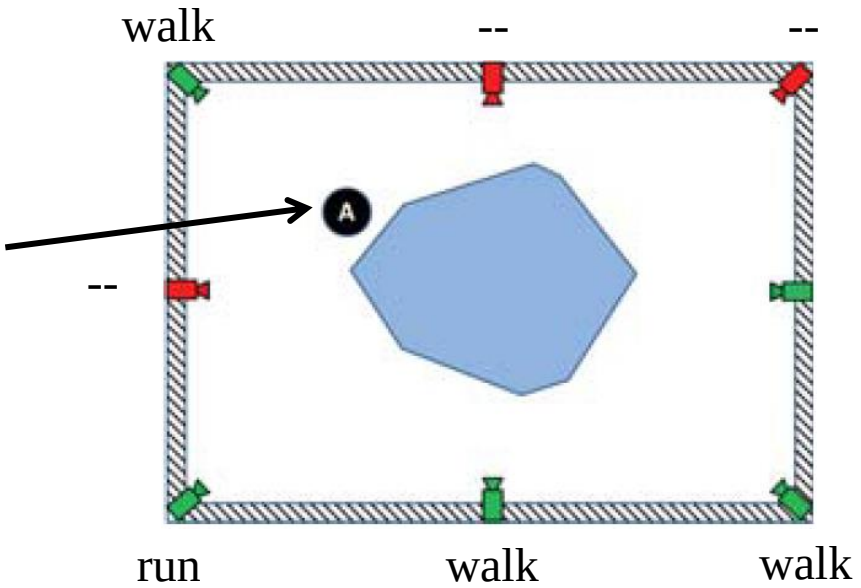


# HAR based on global body information

View-invariant action recognition:

- Exploit one or multiple cameras during training in order to learn an action description which is invariant to the viewing angle
- During the online process, one camera can be used for the recognition of actions (multiple cameras can be used for fusing information and enhancing performance)

A person performing an action captured by $N \leq N_C$ cameras results to the creation of N test action vectors $\mathbf{s}_{test,i}$.



# HAR based on global body information

View-invariant action recognition:

- Train a view-independent model using all available views

Cluster all human body poses obtained for the training videos, to define human body pose prototypes

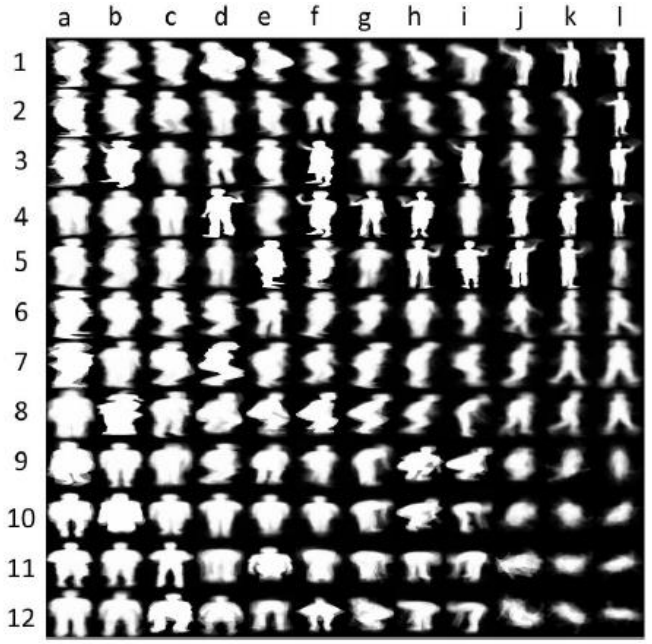
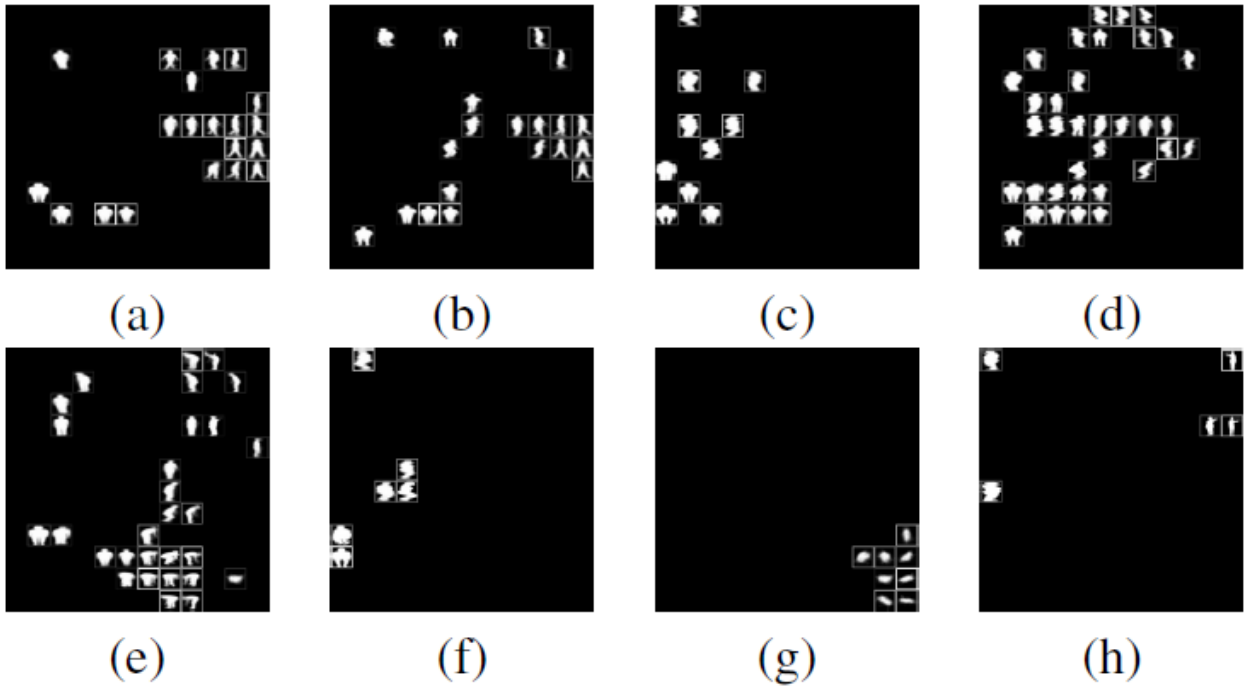



Fig. 3. *A 12 × 12 SOM produced by posture frames of eight actions captured from eight viewing angles.*

# HAR based on global body information

View-invariant action recognition:

- Train a view-independent model using all available views

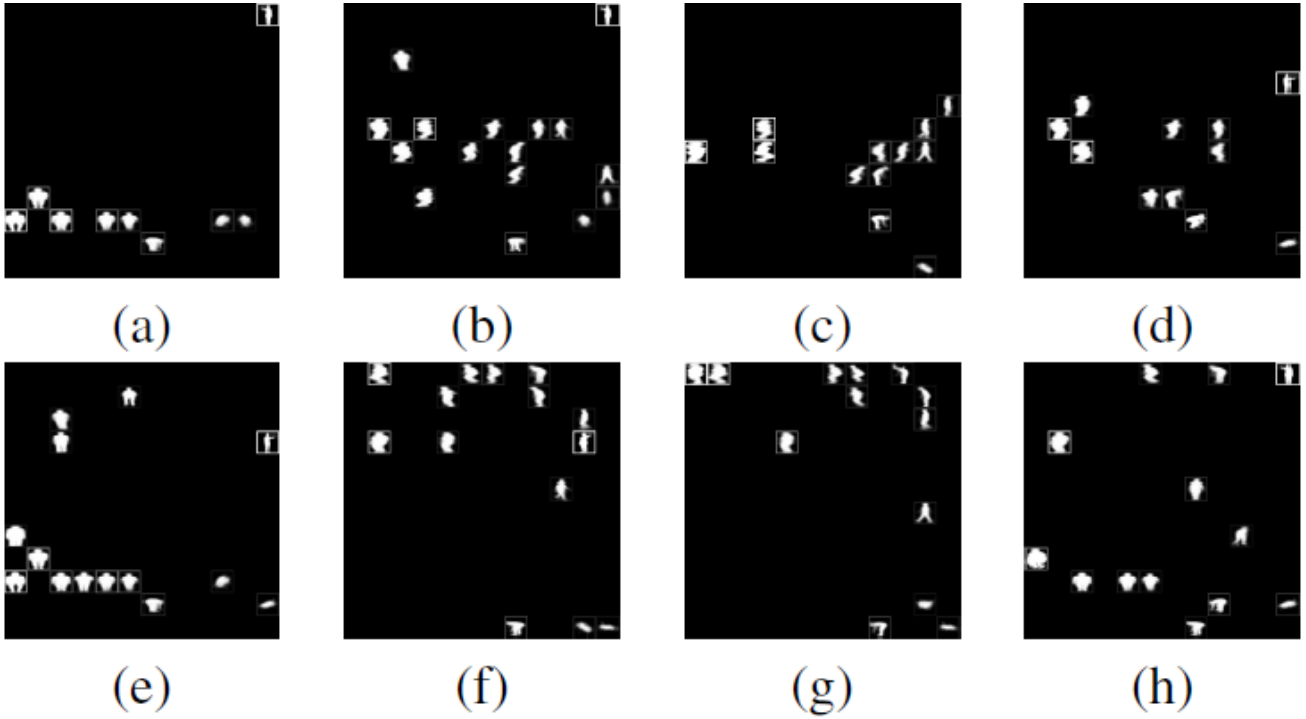



Prototypes corresponding to actions: a) walk, b) run, c) jump in place, d) jump forward, e. bend, f) sit, g) fall and h) wave one hand

# HAR based on global body information

View-invariant action recognition:

- Train a view-independent model using all available views




Prototypes corresponding to views: a) 0°, b) 45°, c) 90°, d) 135°, e) 180°, f) 225°, g) 270° and h) 315°

# HAR based on global body information

View-invariant action recognition:

- Train a view-independent model using all available views

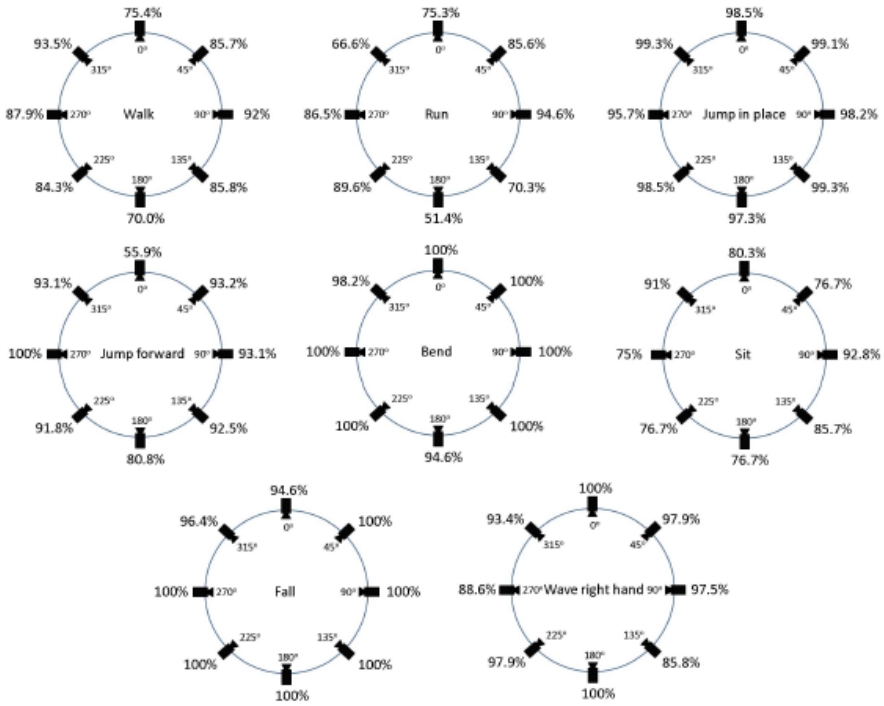
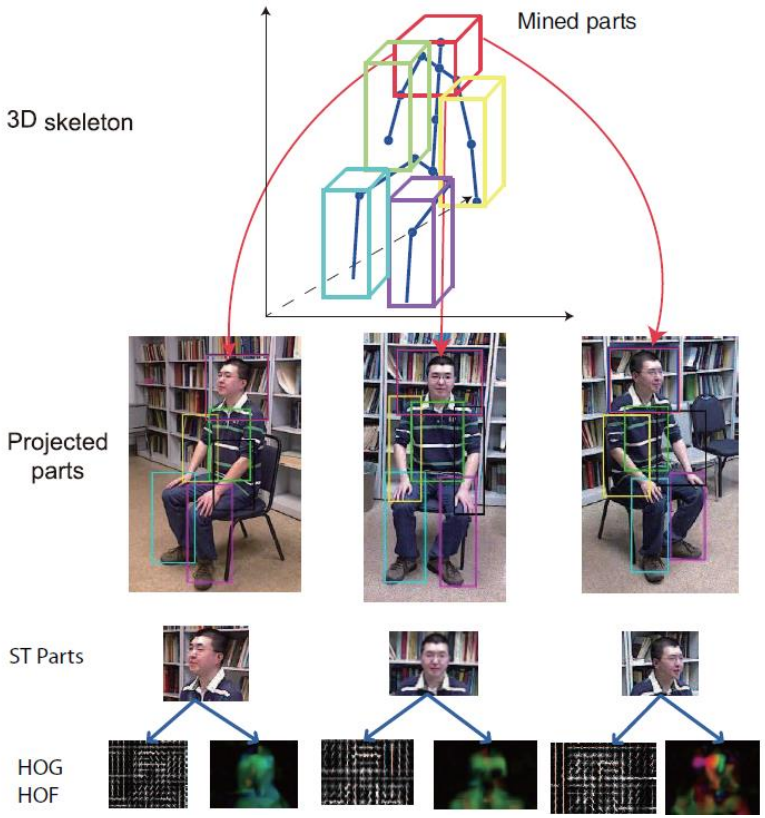



Fig. 6. *Single-view action classification results presented as input to the Bayesian framework for eight actions captured from eight viewing angles.*

# HAR based on global body information

View-invariant action recognition:

- Define a mapping from any possible view to a “reference view”
- This approach leads to cross-view and view-invariant recognition

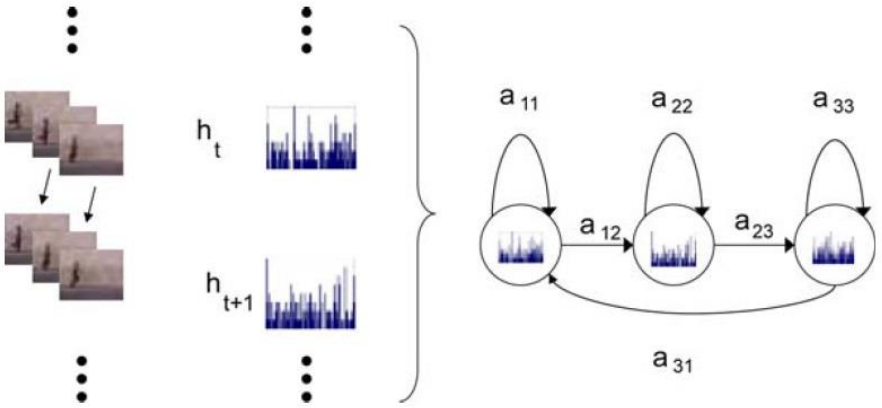


# HAR based on global body information

Action recognition steps:

- Human body poses calculation (action description)
- Action representation calculation (MEI, MHI, Hist, MVs, MvAIs, etc.)
- Action classification:
  - Use of a standard classification method, e.g. k-NN, SVM, LDA+NC, ANN.
  - Use of a generative model that can exploit the sequence of human body poses (e.g. Hidden Markov Model). During evaluation, a video is compared with all action models and is classified to the one providing the maximal score

HMM description



# HAR based on local video information

Methods exploiting global human body information set the assumptions:

- There is one person in the video
- His/hers silhouettes can be calculated in a robust way

These assumptions are unrealistic in many cases!

# HAR based on local video information

Methods exploiting global human body information set the assumptions:

- There is one person in the video
- His/hers silhouettes can be calculated in a robust way

These assumptions are unrealistic in many cases!

But not in all cases!


# HAR based on local video information

In order to describe actions:

- Detect points of interest (video locations that are probable to correspond to actions)
- Describe the shape and motion of a small neighborhood around these interest points.

Space-Time Interest Points (STIPs)

# HAR based on local video information

Interest Points (IPs) in images (Harris detector):

- Given an image $f^{sp}$, detect locations with image values (intensities) that are undergo significant variation in both x and y directions:
- For a given scale $\sigma_l^2$, such IPs can be found by (setting $\sigma_i^2 = \sigma_l^2$):

$$\mu^{sp} = g^{sp}(\cdot;\, \sigma_i^2) * \begin{pmatrix} (L_x^{sp})^2 & L_x^{sp} L_y^{sp} \\ L_x^{sp} L_y^{sp} & (L_y^{sp})^2 \end{pmatrix}$$

where $L_x^{sp}$ and $L_y^{sp}$ are Gaussian derivatives and $g^{sp}$ is the Gaussian kernel:

$$L_x^{sp}(\cdot;\, \sigma_l^2) = \partial_x (g^{sp}(\cdot;\, \sigma_l^2) * f^{sp})$$

$$L_y^{sp}(\cdot;\, \sigma_l^2) = \partial_y (g^{sp}(\cdot;\, \sigma_l^2) * f^{sp})$$

$$g^{sp}(x, y;\, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp(-(x^2 + y^2)/2\sigma^2)$$

IPs are detected on the locations corresponding to positive maxima of ($\lambda_1 \leq \lambda_2$ are the eigenvalue of $\mu^{sp}$):

$$H^{sp} = \det(\mu^{sp}) - k\,\mathrm{trace}^2(\mu^{sp}) = \lambda_1 \lambda_2 - k(\lambda_1 + \lambda_2)^2$$

# HAR based on local video information

Space-Time Interest Points (STIPs) in videos are an extension of IPs:

$$\mu = g(\cdot;\, \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}$$

where $\sigma_i^2 = \sigma_l^2$ and $\tau_i^2 = \tau_l^2$.

$$L(\cdot;\, \sigma_l^2, \tau_l^2) = g(\cdot;\, \sigma_l^2, \tau_l^2) * f(\cdot)$$

$$g(x, y, t;\, \sigma_l^2, \tau_l^2) = \frac{\exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/2\tau_l^2)}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}}$$

STIPs are detected on the locations corresponding to positive maxima of ($\lambda_1 \leq \lambda_2 \leq \lambda_3$ are the eigenvalue of $\mu$):

$$H = \det(\mu) - k\,\mathrm{trace}^3(\mu) = \lambda_1 \lambda_2 \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3$$

# HAR based on local video information

Space-Time Interest Points (STIPs) in videos are an extension of IPs:

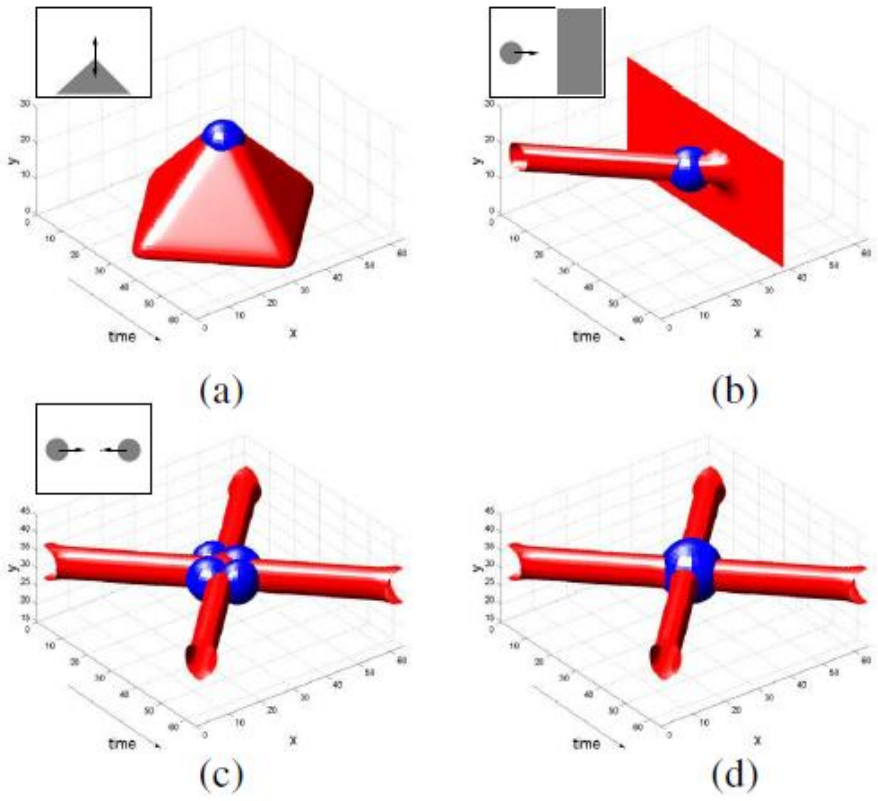


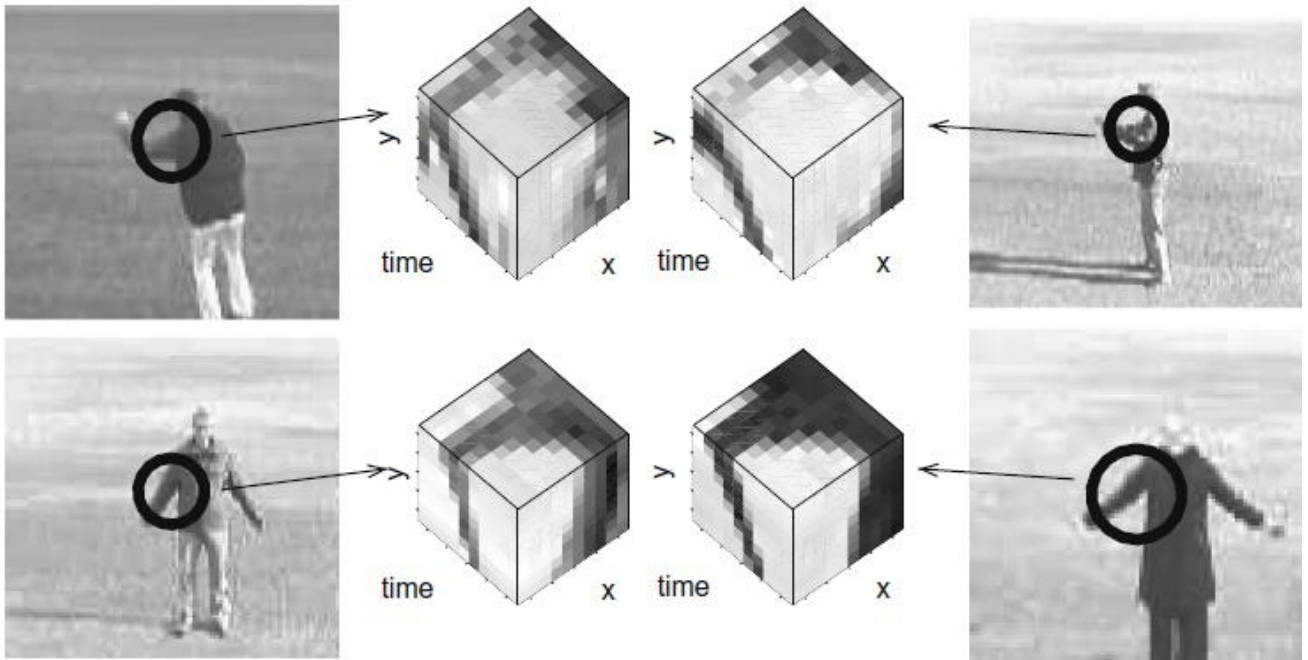

Figure 2: Results of detecting spatio-temporal interest points on synthetic image sequences: (a) Moving corner; (b) A merge of a ball and a wall; (c) Collision of two balls with interest points detected at scales $\sigma_l^2 = 8$ and $\tau_l^2 = 8$; (d) the same as in (c) but with interest points detected at scales $\sigma_l^2 = 16$ and $\tau_l^2 = 16$.

# HAR based on local video information

After the determination of STIPs:

- Description of the shape in the local neighborhood of each STIP (e.g. cuboids, HOG and LBP)

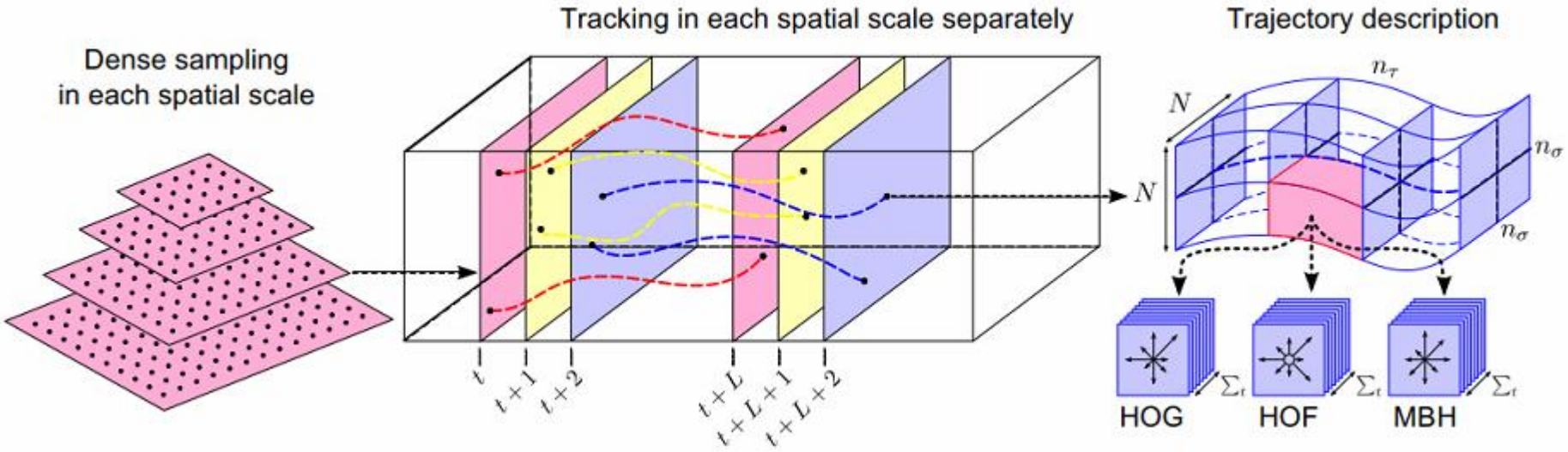
Cuboids



# HAR based on local video information

Dense Trajectories for HAR:

- Interest Point detection in one frame
- IP tracking for L frames (usually L = 14)
- Calculation of local descriptors (HOG, HOF, MBH) on the IP trajectories




Dense Trajectories  Improved DT

# HAR based on local video information

Dense Trajectories for HAR:

- Interest Point detection in one frame
- IP tracking for L frames (usually L = 14)
- Calculation of local descriptors (HOG, HOF, MBH) on the IP trajectories

Dense Trajectories  Improved DT

# HAR based on local video information

Dense Trajectories for HAR:

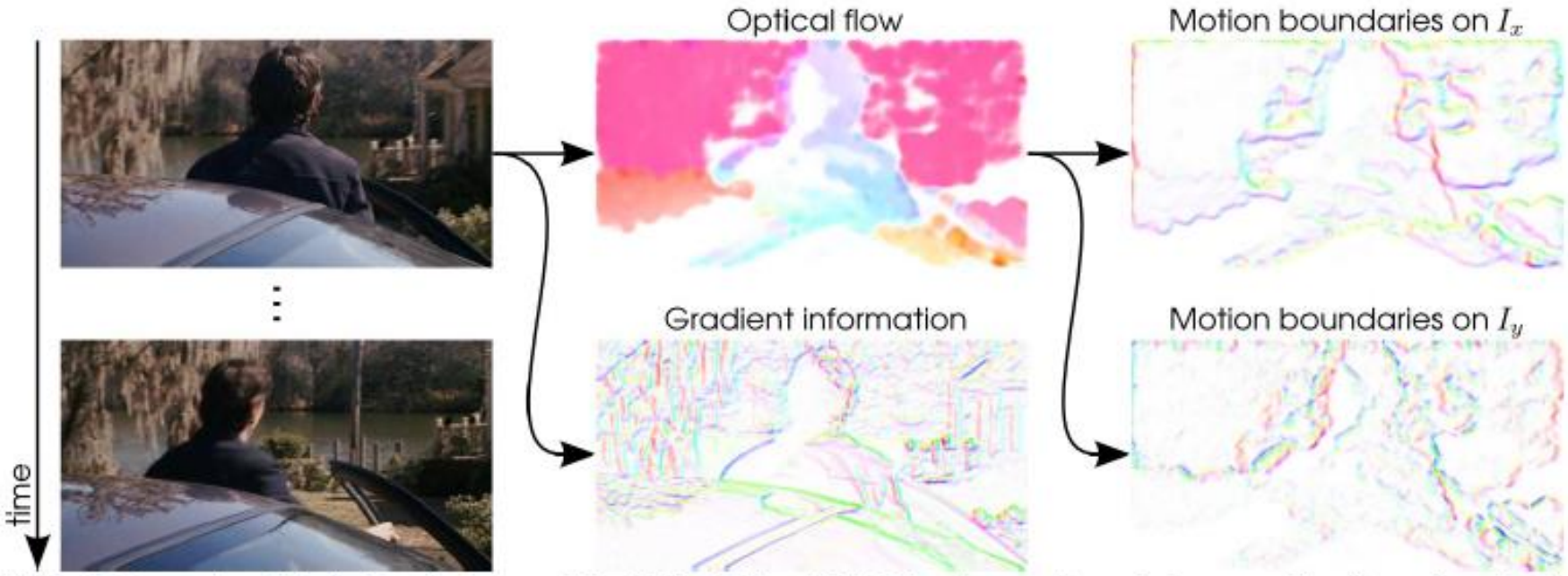


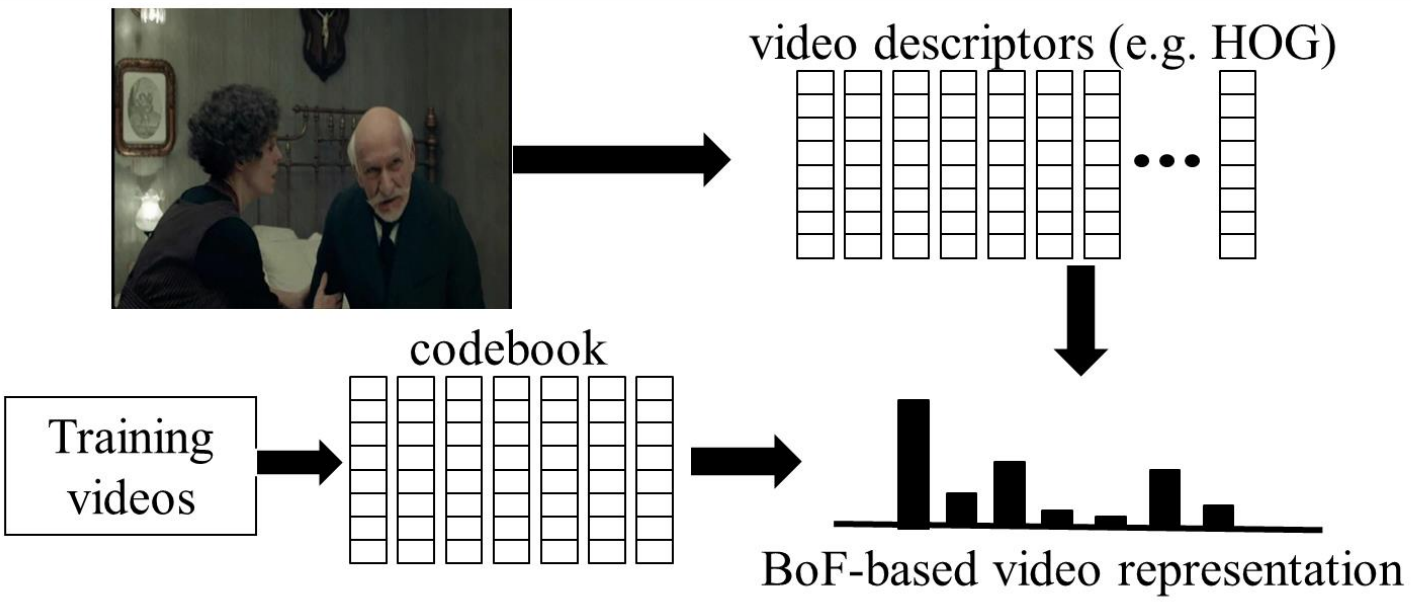

Figure 3. Illustration of the information captured by HOG, HOF, and MBH descriptors. For each image, gradient/flow orientation is indicated by color (hue) and magnitude by saturation. Motion boundaries are computed as gradients of the x and y optical flow components separately. Compared to optical flow, motion boundaries suppress most camera motion in the background and highlight the foreground motion. Unlike gradient information, motion boundaries eliminate most texture information from the static background.

Dense Trajectories  Improved DT

# HAR based on local video information

After the determination of STIPs/DTs:

- Description of the shape in the local neighborhood of each STIP/DT
- Video representation using the BoWs (or BoFs) model:



# HAR based on local video information

After the determination of STIPs/DTs:

- Description of the shape in the local neighborhood of each STIP/DT
- Video representation using the BoWs (or BoFs) model
- Fusion at the classification level using a modified RBF kernel ($D(\cdot, \cdot)$ is the $\chi^2$-distance):

$$K(x_i, x_j) = exp(-\sum_c \frac{1}{A^c} D(x_i^c, x_j^c))$$

- What do we achieve by using STIPs and DTs?
  - We focus the video description in specific locations (which hopefully correspond to the places of actions)
  - We reduce processing time

# HAR based on local video information

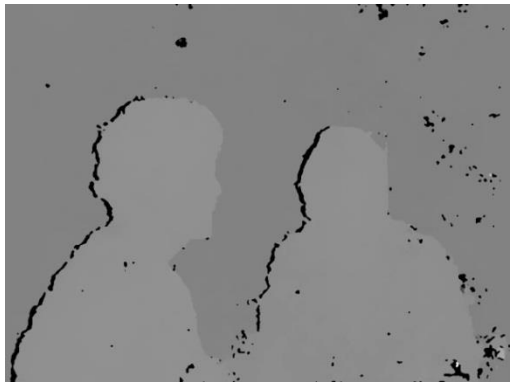
Use of stereo-cameras

- Use the enriched information to change the video description and representation

Left channel

Right channel

Disparity map



Disparity zone 2

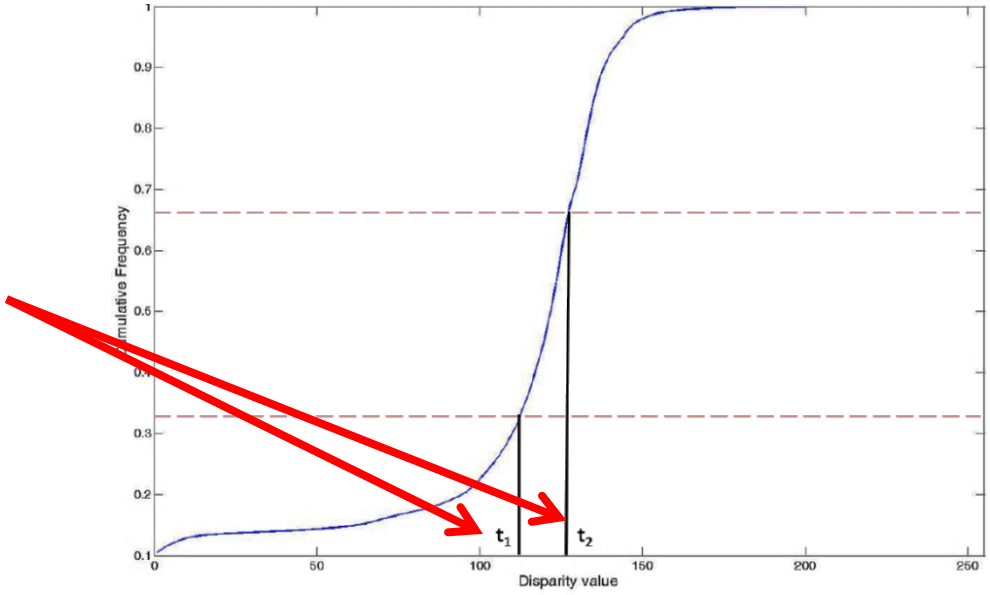
Disparity zone 1

# HAR based on local video information

Use of stereo-cameras

- Use the enriched information to change the video description and representation

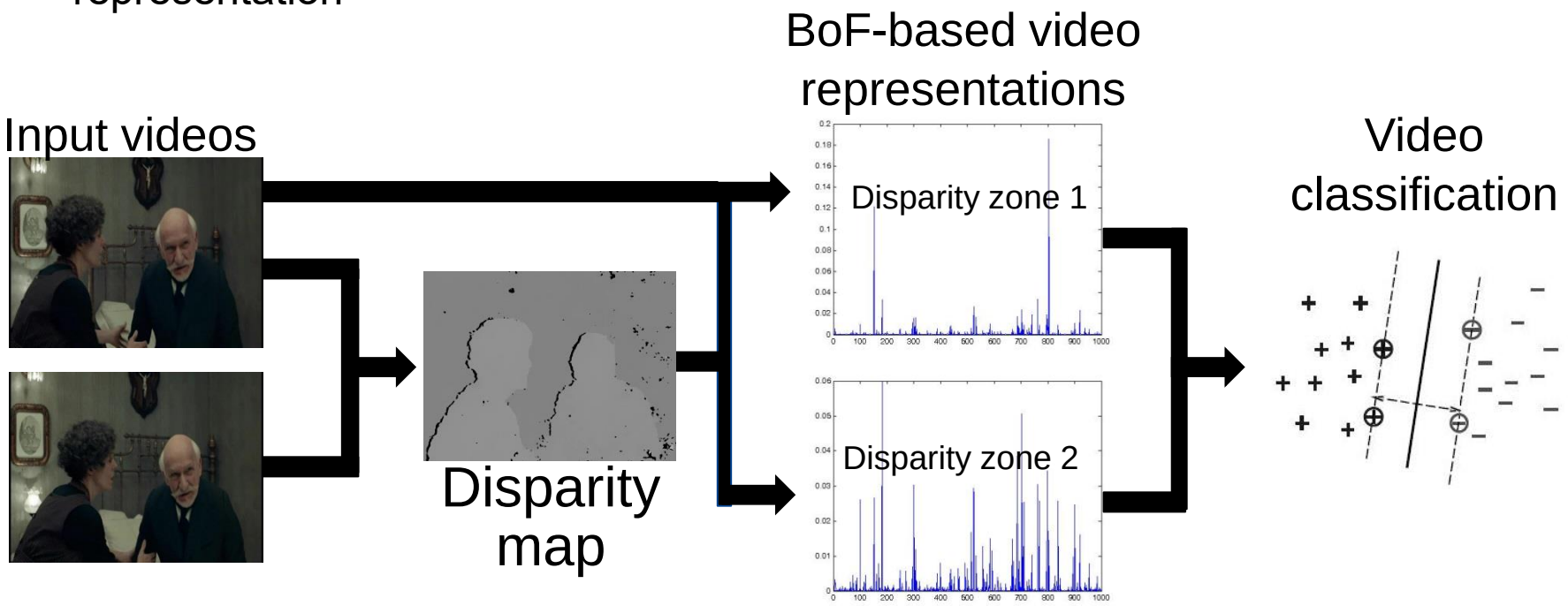
Disparity thresholds



# HAR based on local video information

Use of stereo-cameras

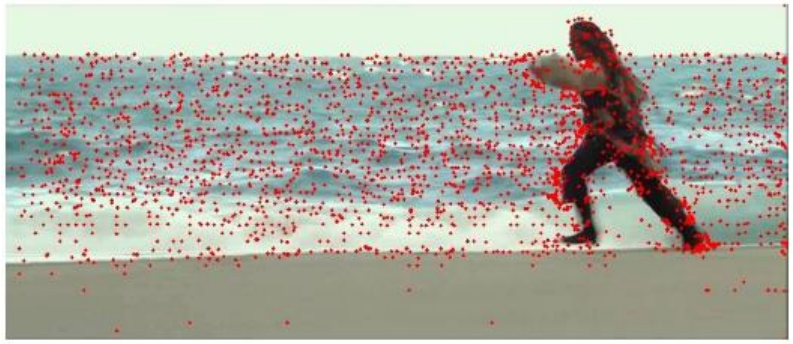
- Use the enriched information to change the video description and representation



# HAR based on local video information

Use of stereo-cameras

- Use the enriched information to change the video description and representation



Original STIPs

Disparity enhanced STIPs

# Neural network-based HAR

Use of Convolutional Neural Networks that can take as input (raw) image/video data for action recognition.

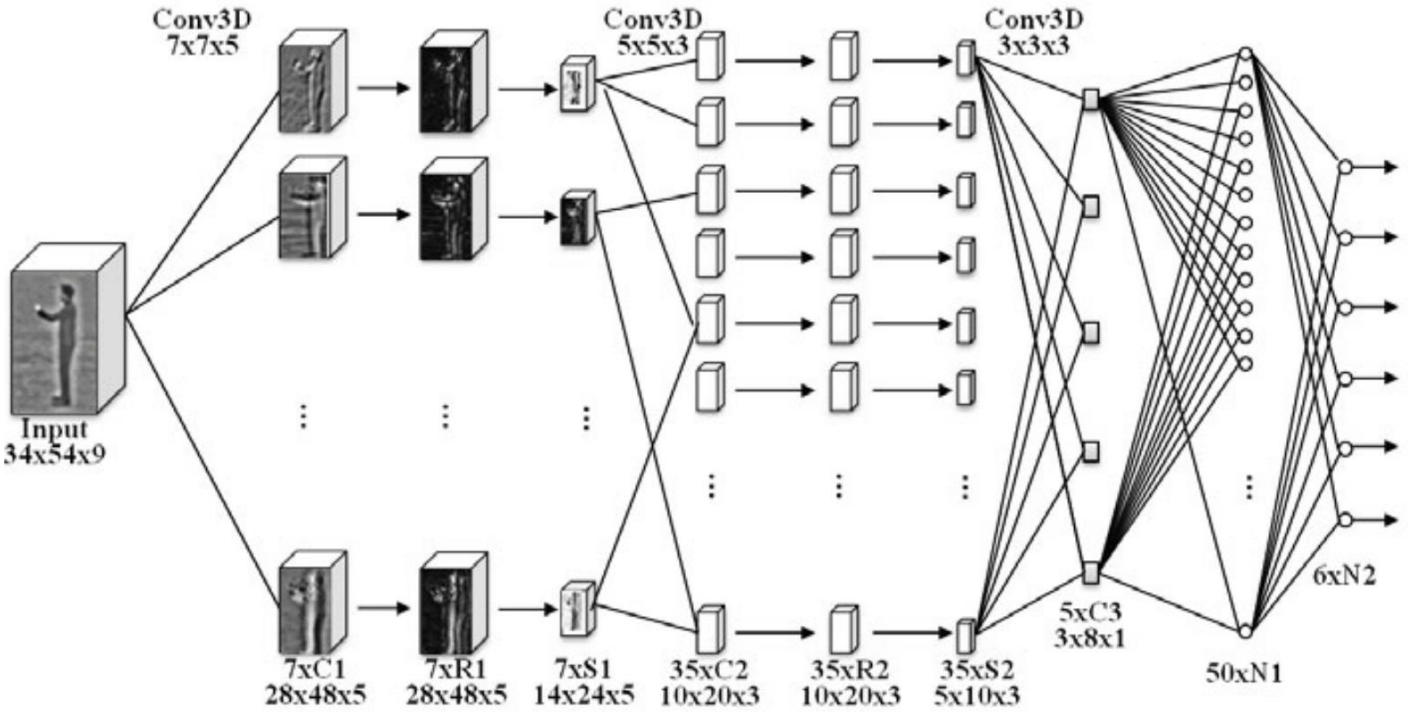
Four types of networks can be used:

- Spatiotemporal networks
- Multiple stream networks
- Deep generative networks
- Temporal coherency networks

# Neural network-based HAR

Spatiotemporal networks

- Use of 3D convolutions in order to exploit space (video frame intensities) and time information in videos



# Neural network-based HAR

Spatiotemporal networks

- Use of 3D convolutions in order to exploit space (video frame intensities) and time information in videos

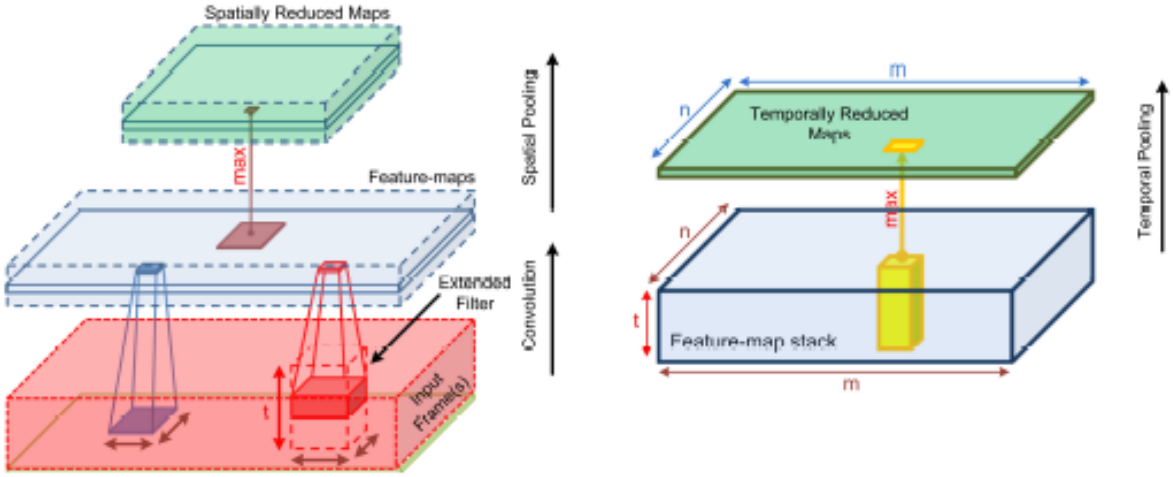



Figure 10: Spatiotemporal operations: 2D convolution (blue), 3D convolution on frame stacks (red) as in Ji et al. (2013), conventional spatial max-pooling (brown), and temporal max-pooling (yellow) as in Ng et al. (2015).

# Neural network-based HAR

In order to exploit the time information we can also use a combination of CNN and Long-Short Term Memory (LSTM) networks

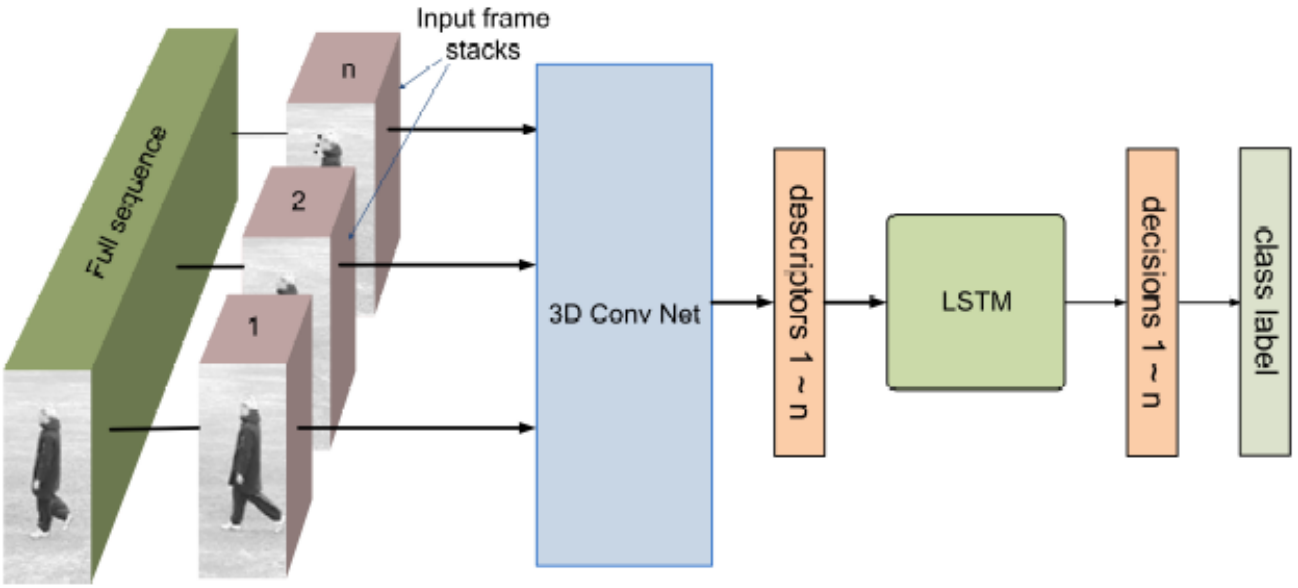



Figure 13: The network structure of (Baccouche et al., 2011).

# Neural network-based HAR

In order to exploit the time information we can also use a combination of CNN and Long-Short Term Memory (LSTM) networks

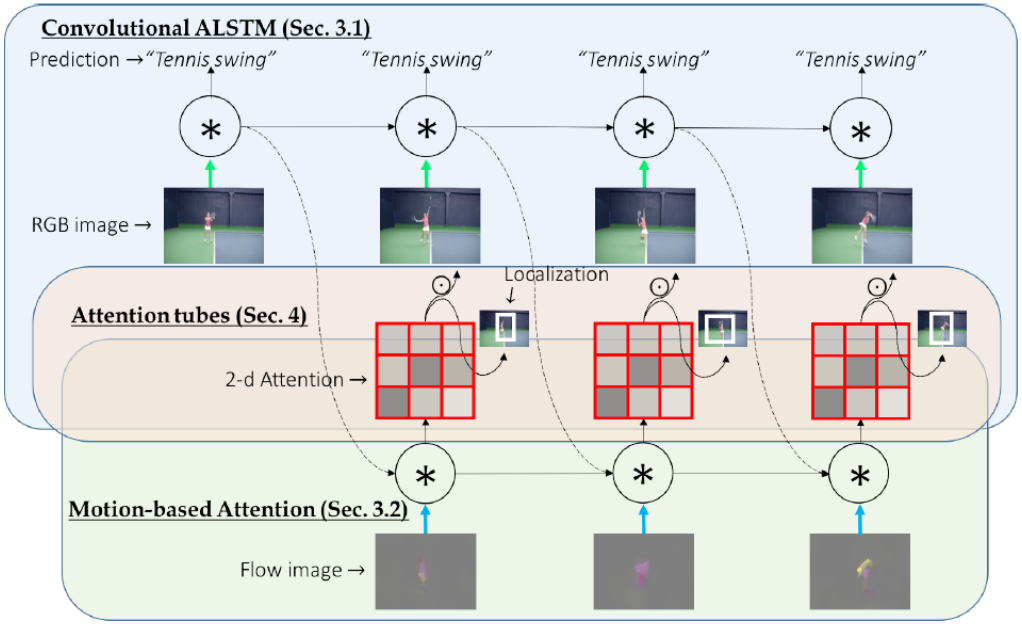



Fig. 1: **The proposed *VideoLSTM* network.** The blue container stands for the *Convolutional ALSTM* (Sec. 3.1), the green container stands for the motion-based attention (Sec. 3.2), while in the pale red container we rely on attention maps for action localization (Sec. 4).

# Neural network-based HAR

Multiple Stream networks

- Except using only the visual information (video frames) multiple types of information (usually visual and optical flow) can be combined.

# Neural network-based HAR

## Multiple Stream networks

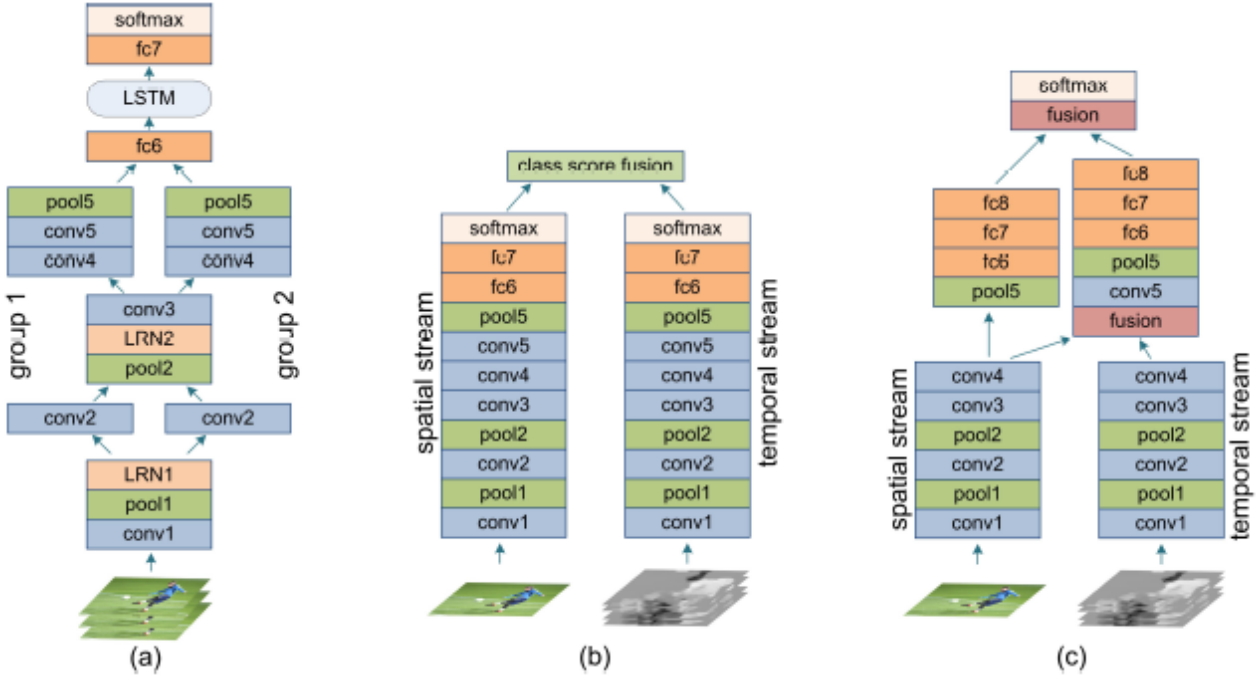



Figure 14: (a) LRCN network structure of (Donahue et al., 2015). A *group* is a set of convolutional filters operating only on a particular set of feature maps from the previous layer. For clarity, we denote each group by a separate convolutional blocks. (b) The two-stream network by Simonyan and Zisserman (2014) with RGB and stacked optical-flow frames as inputs. (c) An example of a two stream fusion network of Feichtenhofer et al. (2016).

# Neural network-based HAR

Deep generative models

- Use of generative models (network architectures) that can predict the future of the video sequence in an unsupervised manner.

# Neural network-based HAR

## LSTM Auto-Encoder Model

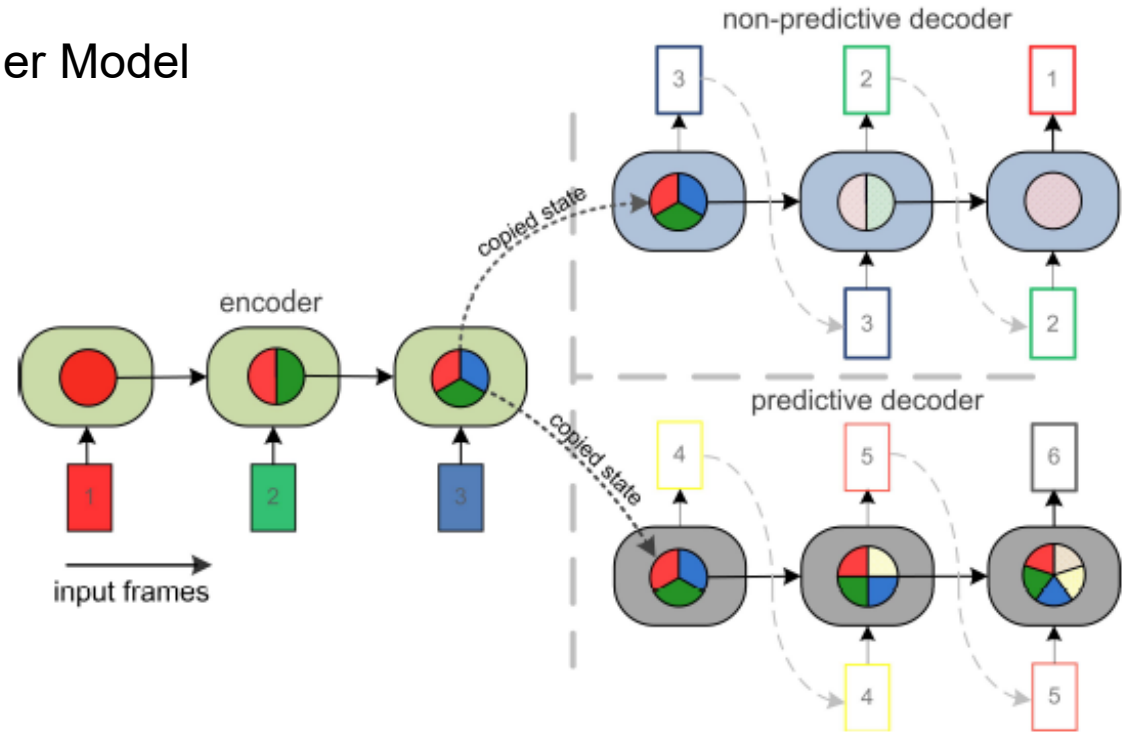



Figure 15: The composite generative LSTM model by Srivastava et al. (2015a). The internal states (represented by the circle inside) of the encoder LSTM captures a compressed version of the input sequence (*e.g.*, frames 1, 2 and 3). The states thereafter are copied into two decoder models, which are reconstructive and predictive. The reconstruction decoder attempts to reconstruct original frames in the reverse order. The predictive model is trained on predicting the future frames 4, 5 and 6. The colors on the state markers indicate the presence of information from a particular frame.

# Neural network-based HAR

Temporal coherency networks

- Assumption: consecutive video frames are correlated both semantically and dynamically

# Neural network-based HAR

Siamese triplet network trained for both pose and action classification

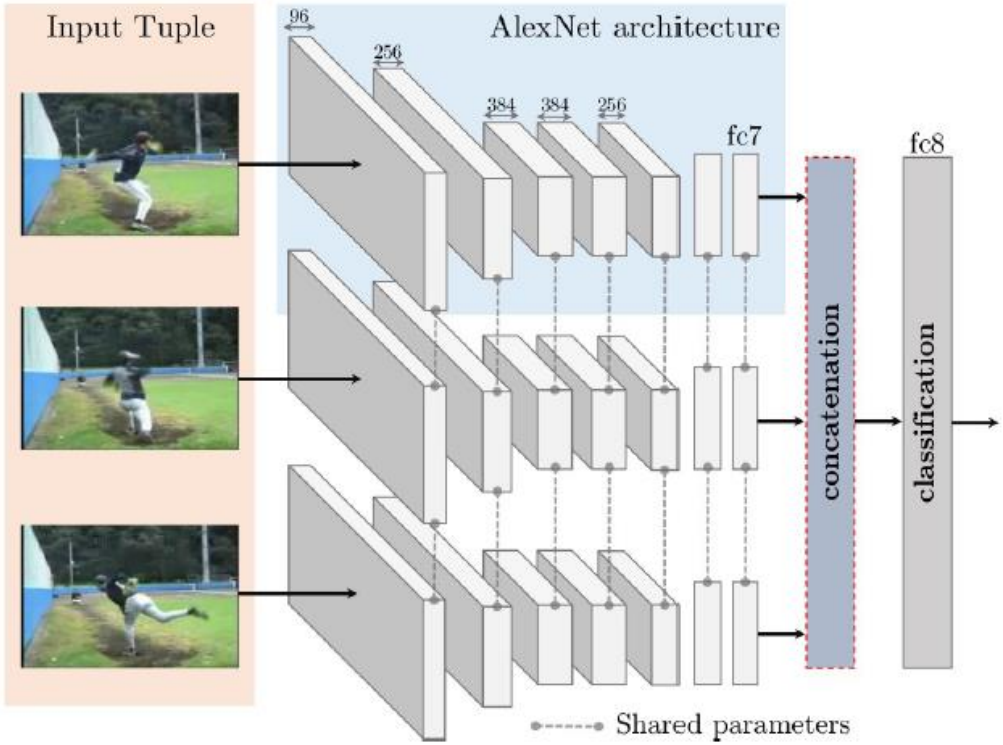



Figure 16: The Siamese Triplet network used by Misra et al. (2016). Each network is expected to capture the motion and pose of actions.

# Neural network-based HAR

Split of action in two phases (Precondition set $X_p$ and Effect set $X_e$)

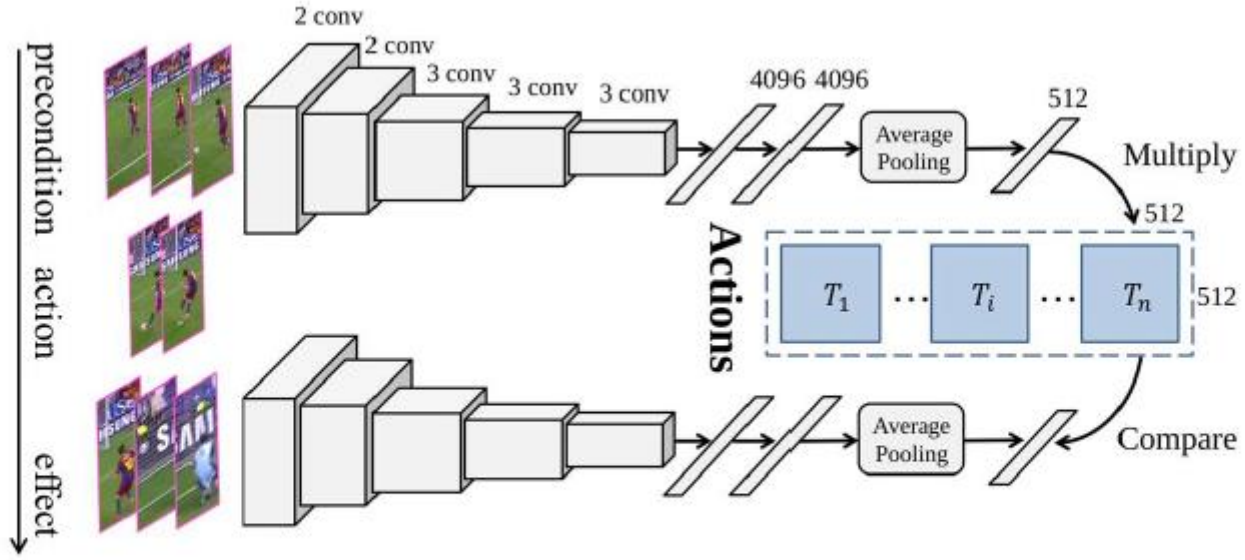



Figure 17: The parallel convolutional structures are used in extraction of precondition and post-effect features.