Action Recognition in Low Quality Videos by Jointly Using Shape, Motion and Texture Features

Saimunur Rahman, John See and Ho Chiung Ching

Center of Visual Computing

Multimedia University, Cyberjaya

CENTRE OF VISUAL COMPUTING

MMU MULTIMEDIA UNIVERSITY

Motivation

- Local space-time features have become popular for action recognition in videos.
- Current methods focus on *high quality videos* which are not suitable for real-time video processing applications.
- Current methods handles various complex video problems (such as *camera motion*) but problem of *video quality* is still relatively unexplored [Oh et al'11].

Goal of this work

- Investigate and analyze the performance of action recognition under two low quality conditions:
  - Spatial downsampling
  - Temporal downsampling
- Joint utilization of shape, motion and texture features for robust recognition of actions from *downsampled* videos.
- Investigate ‘good’ feature combinations for action recognition in low quality video.

Related Works

- Shape and motion features
  - Space-time interest points [Laptev'05]
  - Dense Trajectories [Wang et al.'11]
- Textural features
  - Local Binary Pattern on three orthogonal planes [Kellkompu et al.'08]
  - Extended Local binary pattern on three orthogonal planes [Mattivi and Shao'09]

Outline

- Spatio-temporal video features
- Action recognition framework
- Video downsampling
- Experiments

Spatio-temporal video features

Action recognition framework

Video downsampling

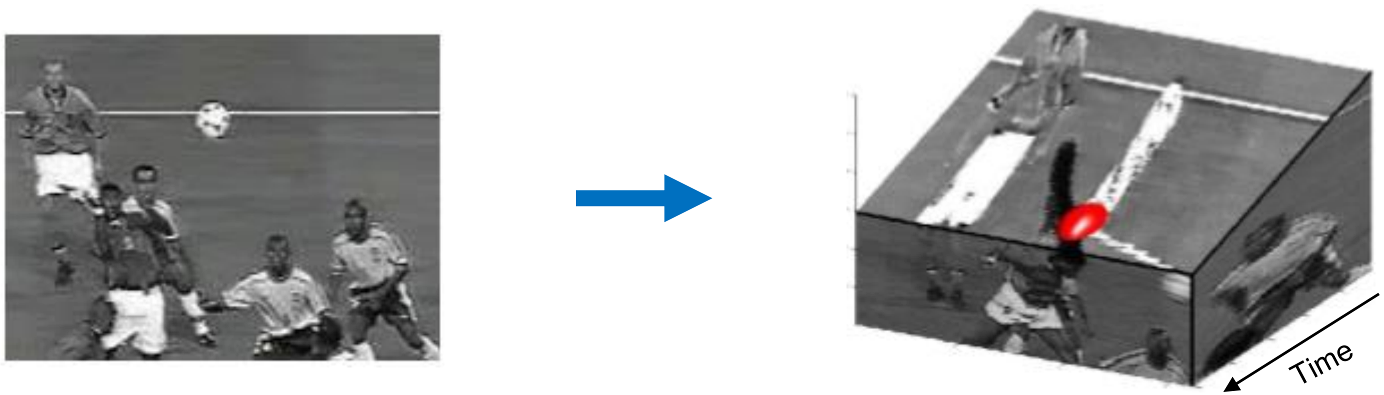
Experiments

Spatio-temporal video features

- Shape and Motion Features *(structures and its change with time)*
  - Feature detector – Harris3D
  - Feature descriptor – HOG and HOF
- Textural Features *(change of statistical regularity with time)*
  - Feature detector and descriptor – LBP-TOP

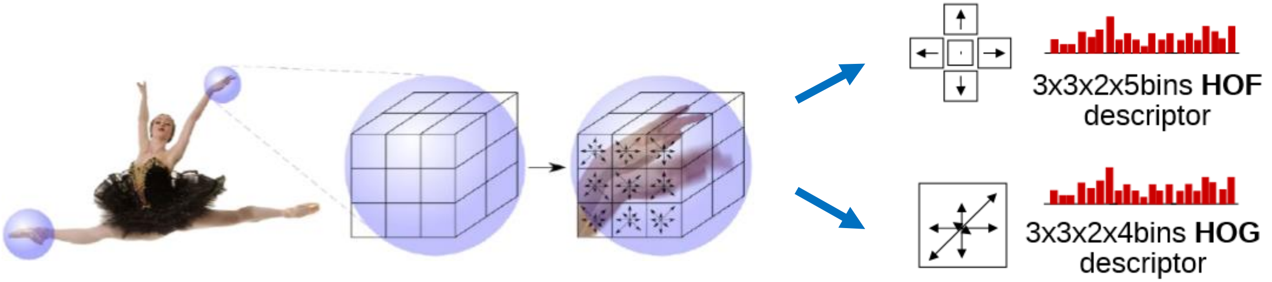
Harris3D detector [Laptev'05]

- Space-time corner detector
- Capable of detecting any spatial and temporal interest point
- Dense scale sampling (no explicit scale selection)



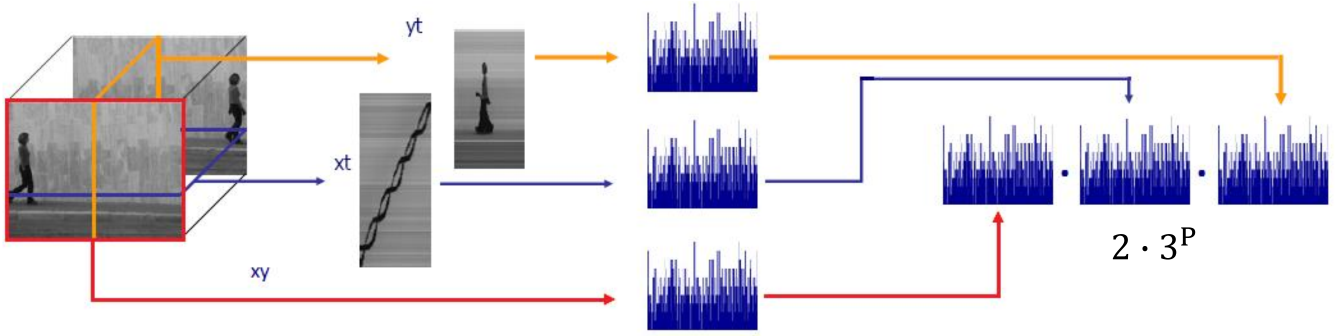
HOG/HOF descriptor [Laptev'08]

- Based on gradient and optical flow information
  - HOG – Histogram of oriented gradients
  - HOF – Histogram of Optical Flow
- Detected 3D patch (xyt) is divided into grid of cells
- Each cell is described with HOG and HOF.



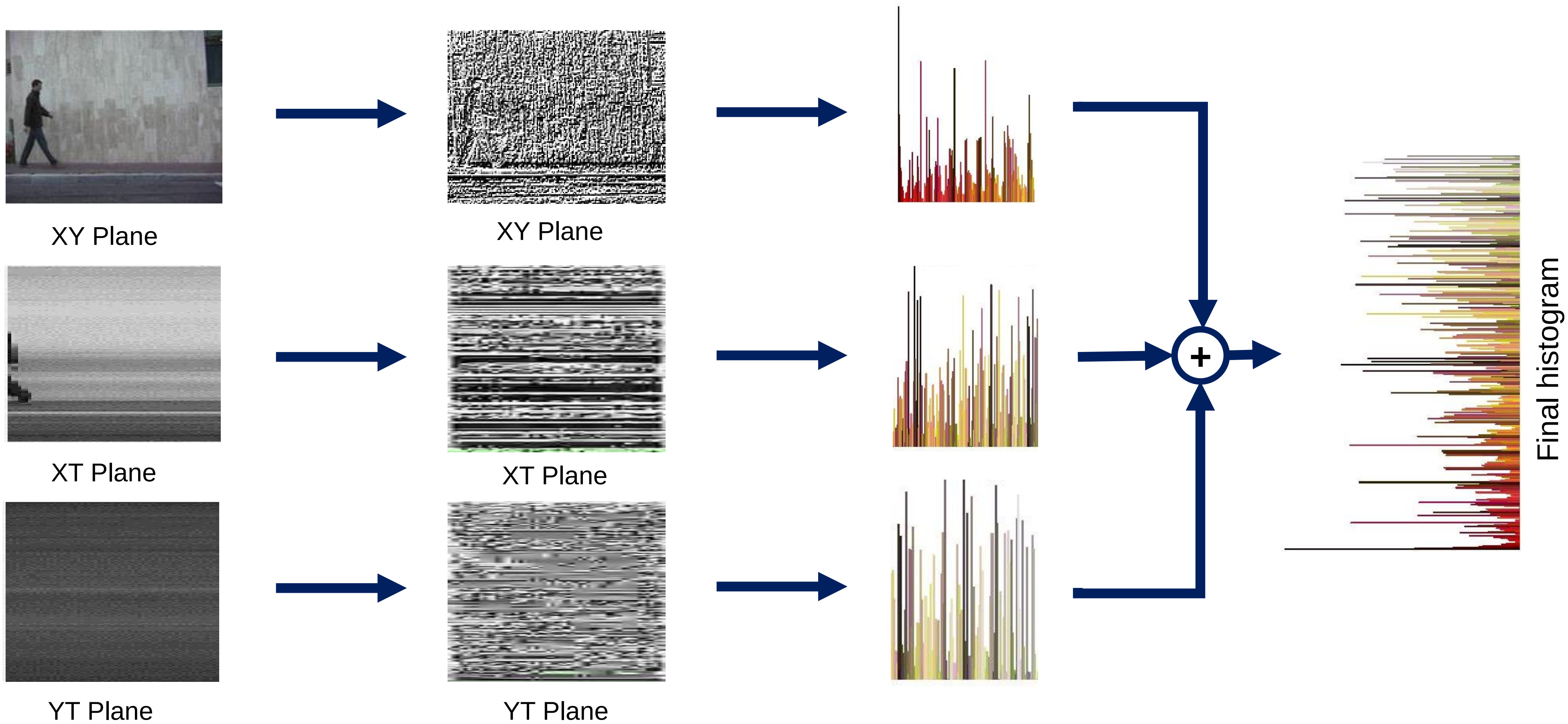
LBP-TOP detector + descriptor [Zhao'07]

- Extension of popular local binary pattern (LBP) operator into three orthogonal planes (TOP)
- Encodes shape and motion on three orthogonal planes (XY, XT and YT)
- Calculate occurrence of different plane histograms to form final histogram ($H = h^{XY} \cdot h^{XT} \cdot h^{YT}$)



$\mathrm{LBP-TOP}_{P_{XY}P_{XT}P_{YT}R_XR_YR_T}$

LBP-TOP in action



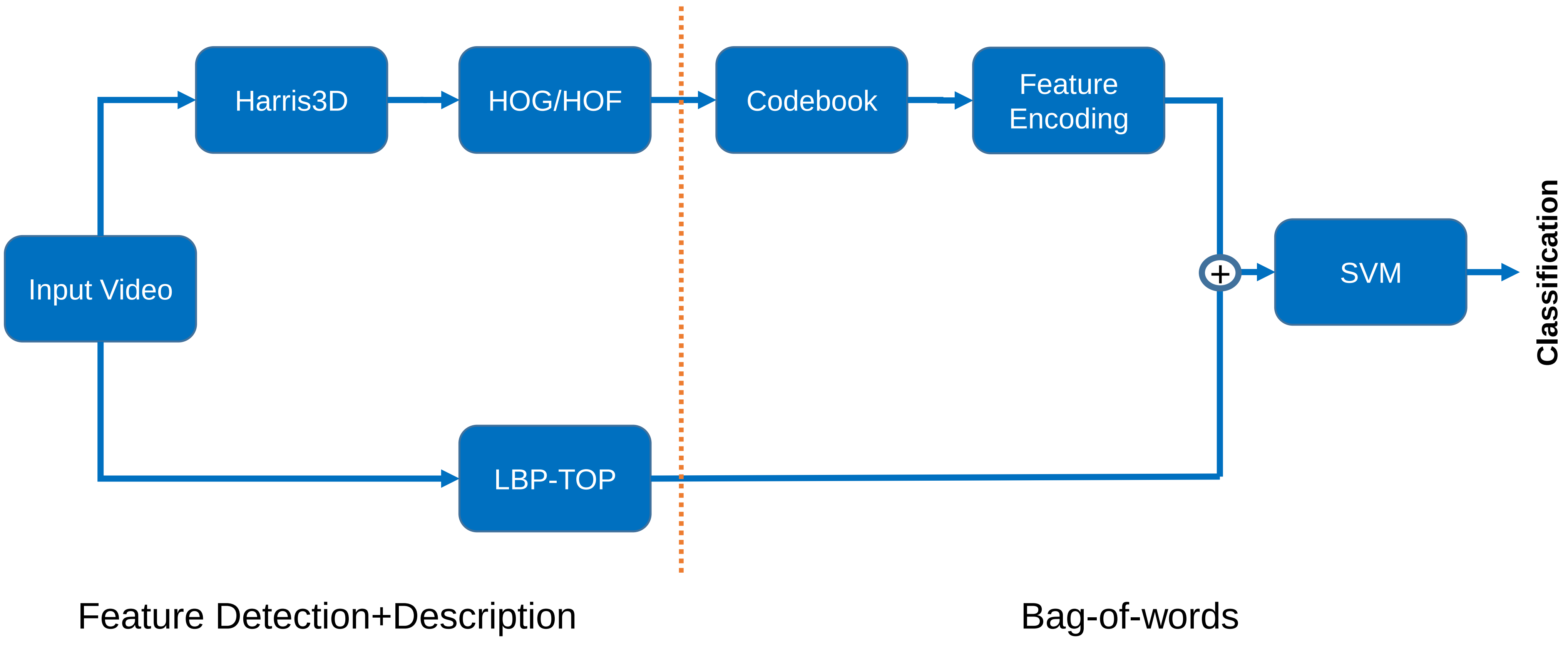
Spatio-temporal video features

Action recognition framework

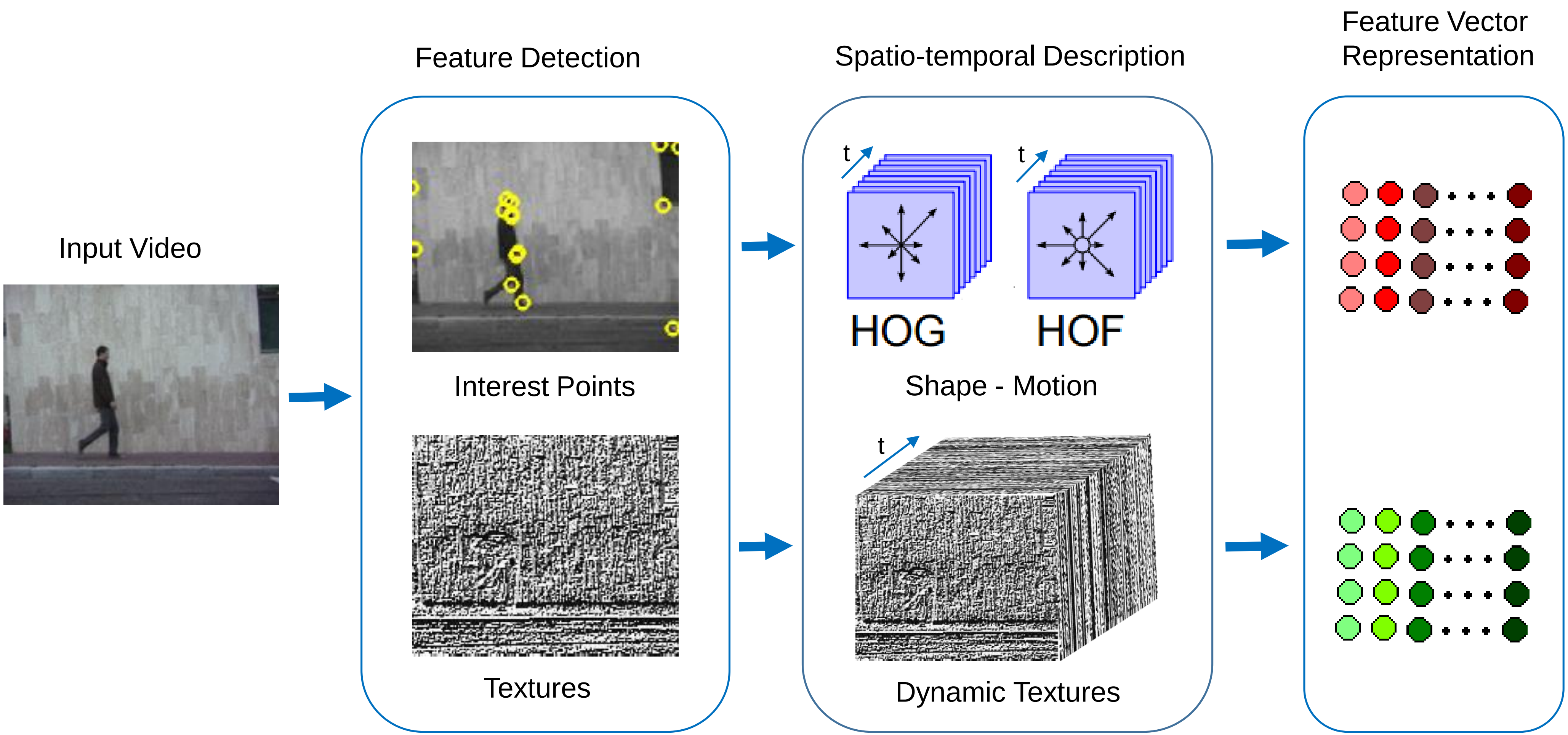
Video downsampling

Experiments

Evaluation framework



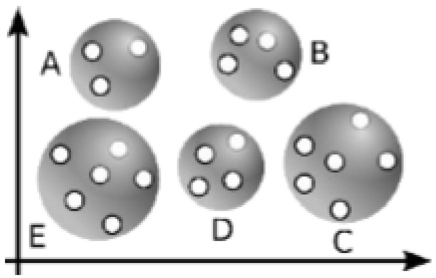
Detection + description of features

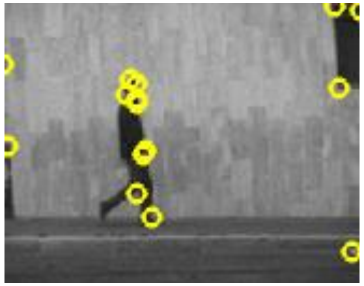


Bag-of-words representation

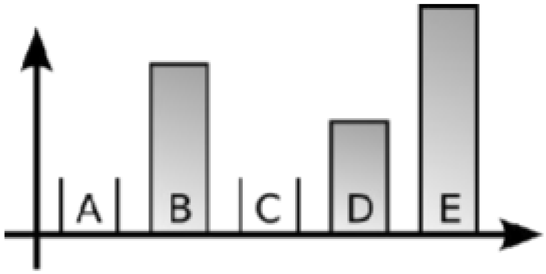
Bag of space-time features + SVM with χ^2 kernel [Vedaldi'08]

Training feature vectors are clustered with k-means





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



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

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

Spatio-temporal video features

Action recognition framework

Video downsampling

Experiments

Video Downsampling

- Spatial downsampling (SD) decrease the spatial resolution.
- Temporal downsampling (TD) reduces temporal sampling rate.

SD Factor	Description
SD_1	Original Res.
SD_2	$^1/_2$ Res. of Original
SD_3	$^1/_3$ Res. of Original
SD_4	$^1/_4$ Res. of Original

TD Factor	Description
TD_1	Original F.R.
TD_2	$^1/_2$ F.R. of Original
TD_3	$^1/_3$ F.R. of Original
TD_4	$^1/_4$ F.R. of Original

Fig: Spatially downsampled videos. (a) SD_1 (b) SD_2 (c) SD_3 (d) SD_4 .

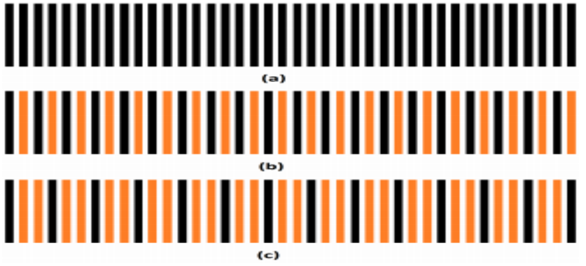
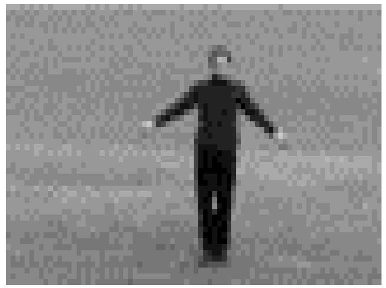


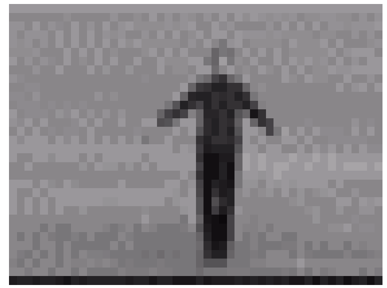
Fig: Temporal Downsampling; (a) Original video (b) TD_2 (c) TD_3

Preview of downsampled videos

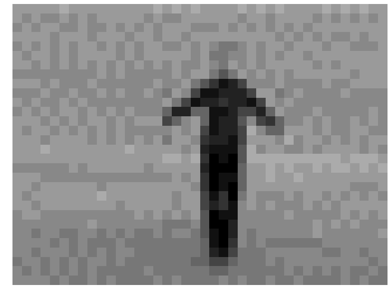
Original Video



SD_2



SD_3



SD_4

TD_2

TD_3

TD_4

Spatio-temporal video features

Action recognition framework

Video downsampling

Experiments

Datasets

- Two popular publicly available dataset
  - KTH action [Schuldt et al.’04]
  - Weizmann [Blank et al.’05]
- Both captured in a controlled environment with homogeneous background.

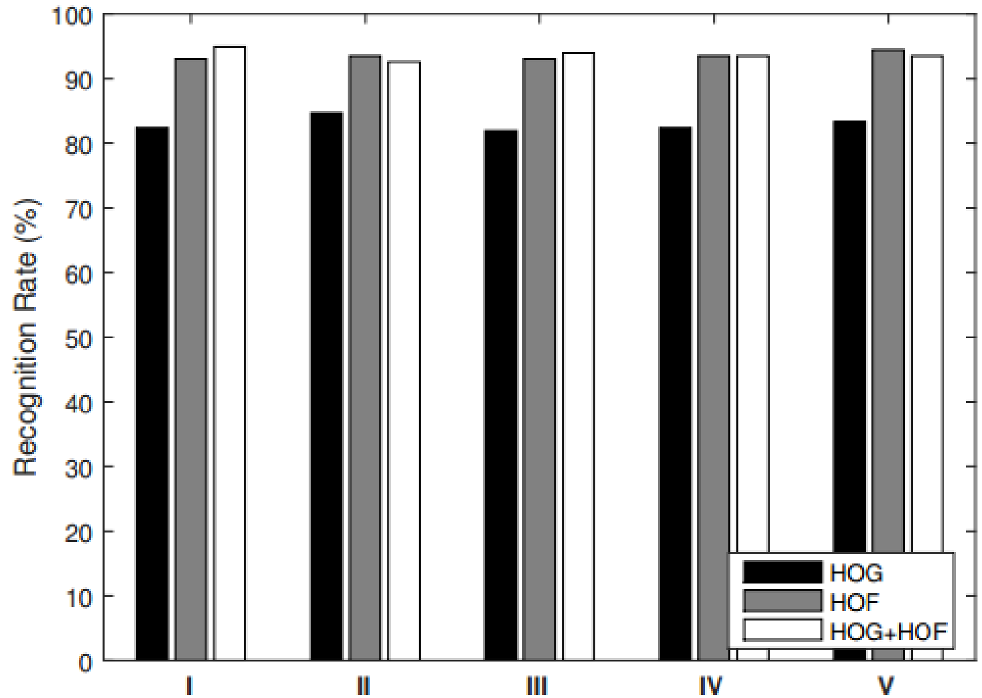
Feature combination used

- Five different feature combinations
  - Combination **I** : (HOG + HOF) - linear kernel
  - Combination **II** : (HOG + HOF) - χ^2 kernel
  - Combination **III** : (HOG + HOF + LBP-TOP) - linear kernel
  - Combination **IV** : (HOG + HOF) + LBP-TOP - χ^2 kernel
  - Combination **V** : (HOG + HOF + LBP-TOP) - χ^2 kernel

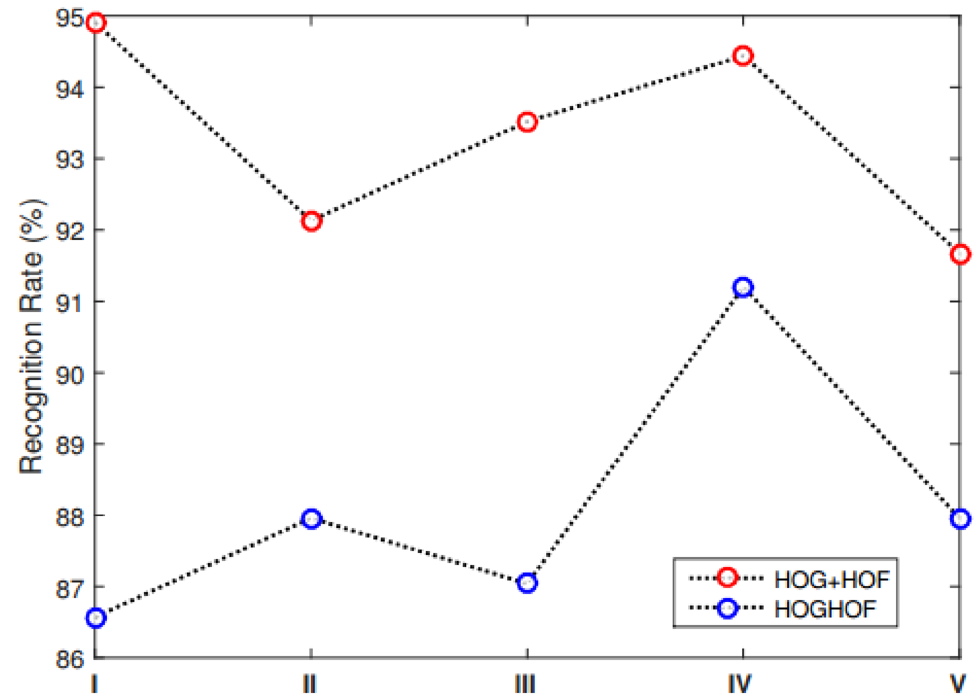
KTH actions [Schuldt et al.'04]

- Total 599 videos divided in 6 action classes
- 25 people performed in 4 different scenarios
- Frame resolution: 160 x 120 pixels
- Frames per second: 25 (average duration 10-15 sec.)
- Followed author specified setup for training-testing splits.
- Performance measure: *average accuracy over all classes*

KTH original dataset - results



KTH dataset

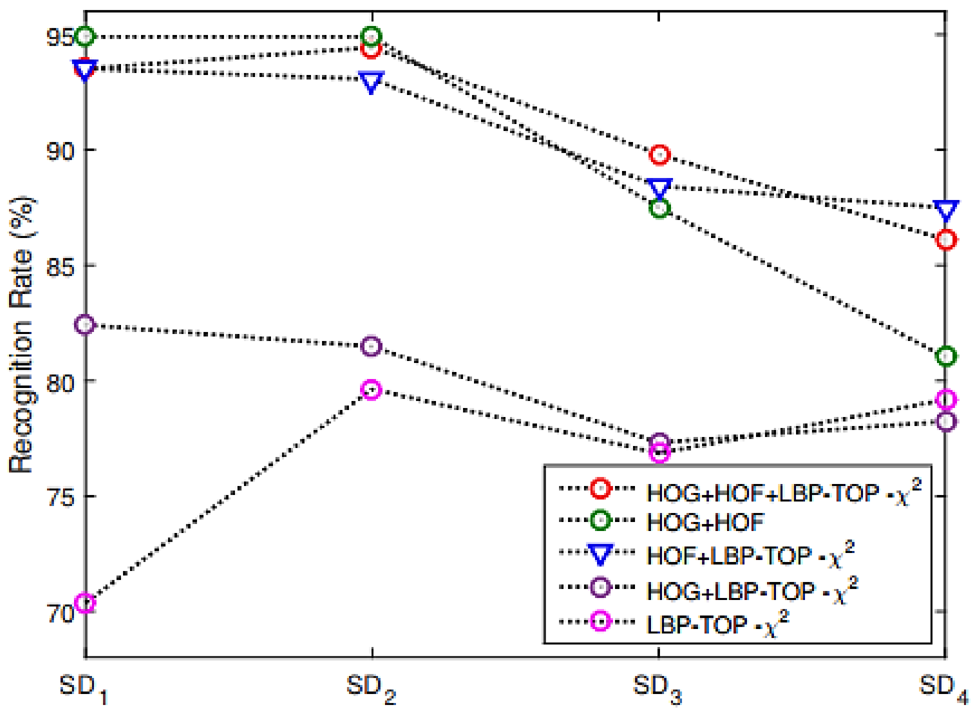


HOGHOF vs. HOG+HOF

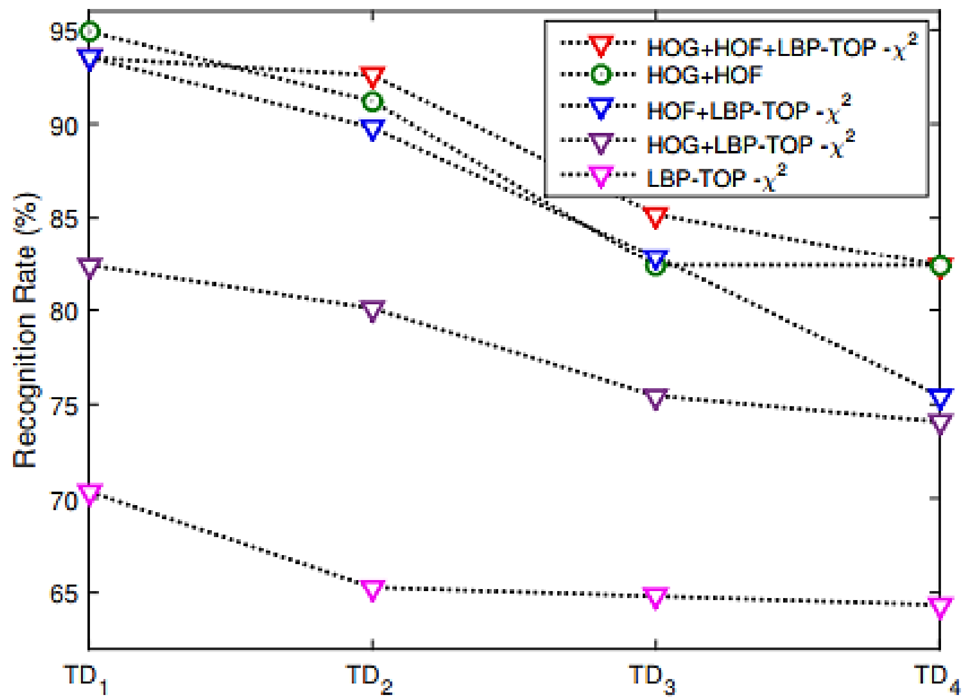
KTH original dataset – results (2)

- Best result for HOG+HOF (94.91%)
- HOG+HOF helps to elevate the overall accuracy by 3–8% ☺
- Kernelization of specific features are able to strengthen results
    - HOF + LBP-TOP : 93.06%
    - HOF + LBP-TOP - χ^2 kernel : 94.44% ☺
- HOF is more effective than HOG but improves when paired with LBP-TOP ☺

KTH downsampled videos – results



Spatial downsampling *(k=2000)*



Temporal downsampling *(k=2000)*

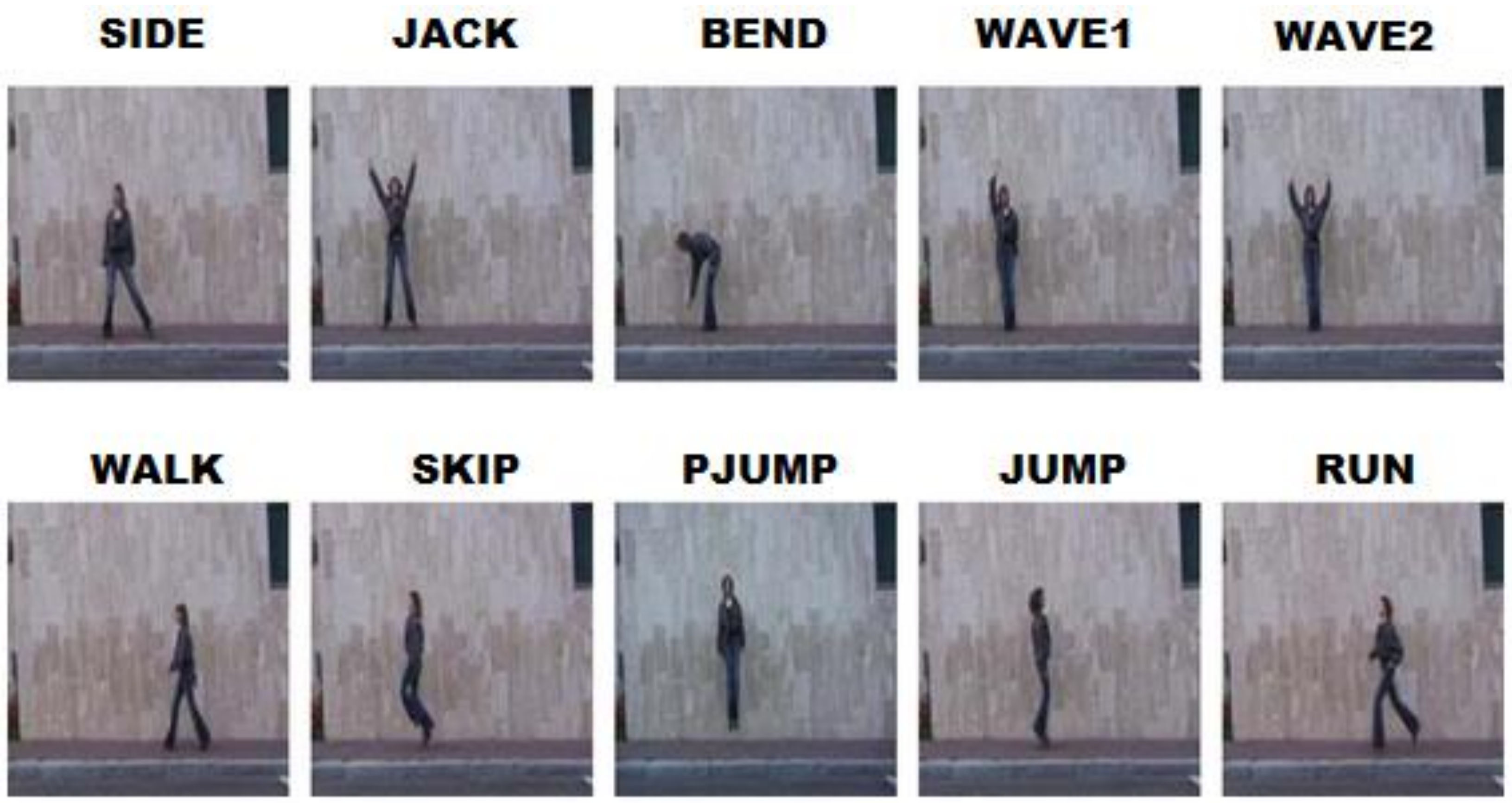
KTH downsampled videos – results (2)

- STIPs and kernalized LBP-TOP appear to dominate the best results within each mode ☺
- LBP-TOP contributes more with the deterioration of spatial or temporal quality (more significant in case of SD_4 & TD_4) ☺
  - Shape information are more important for low temporal resolution ☹
  - Motion information are more important for low spatial resolution ☹
- Note: for STIPs detection in SD modes different *k* parameters are used

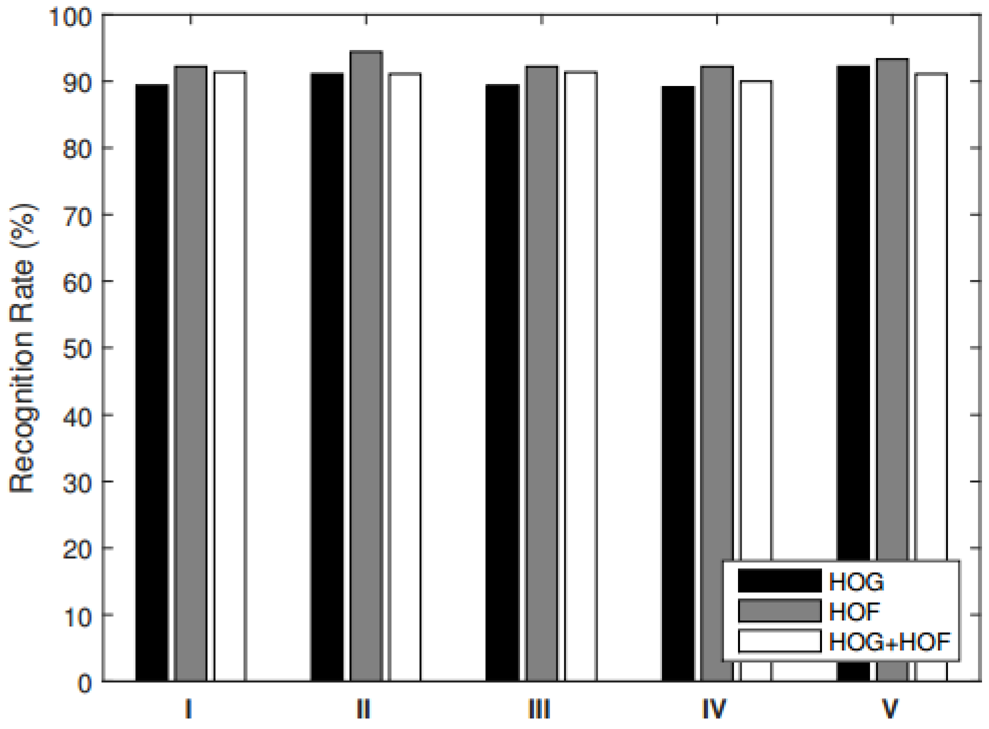
Weizmann [Blank et al'05]

- Total 93 videos divided in 10 action classes
- 9 people performed different actions
- Frame resolution: 180 x 144 pixels
- Frames per second: 50 (average duration 2-3 sec.)
- Performance measure: leave-one-out-cross-validation

Weizmann video sample

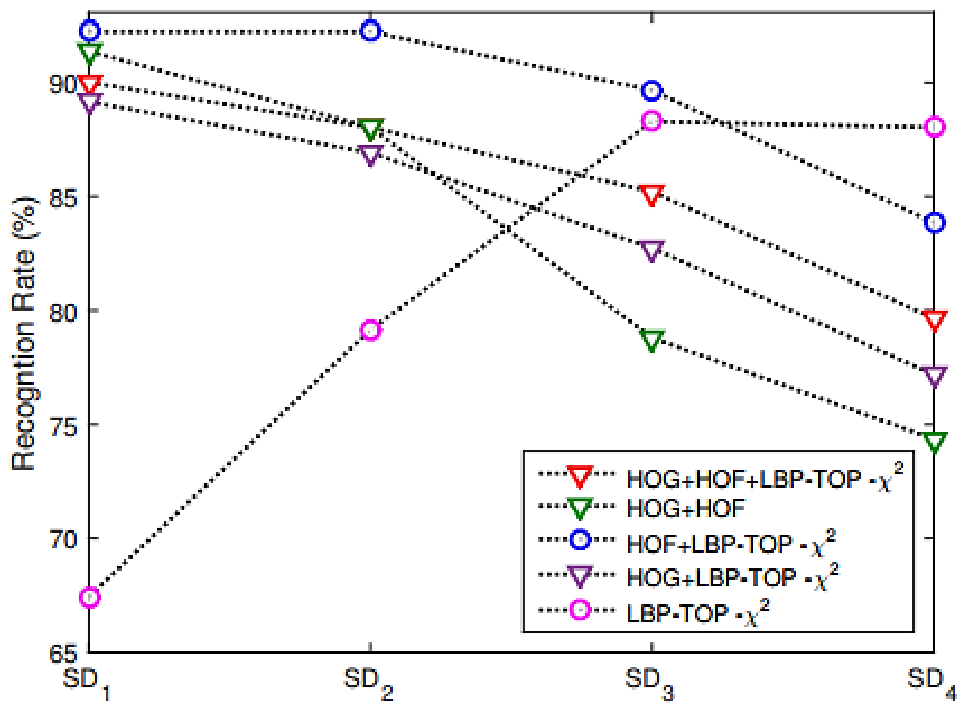


Weizmann original dataset - results

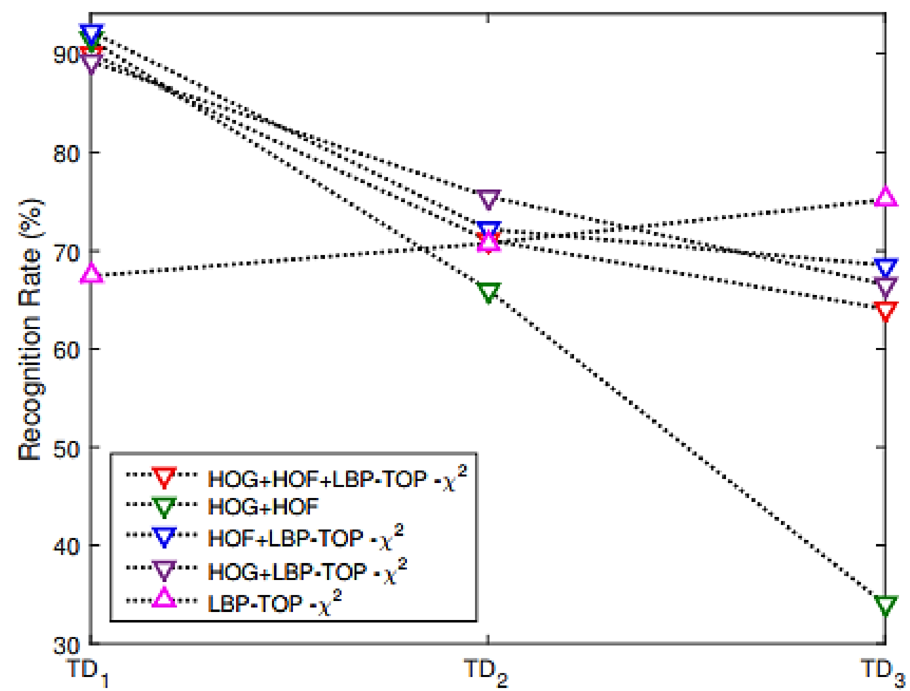


- Best result 94.44% for HOF.
- HOF+LBP-TOP dominate best result within each mode ☺
- Kernelization of LBP-TOP features are able to strengthen results ☺
- Kernelization is less effective for HOF features ☹
- Shape is largely poor on all combinations ☹ but performs better after combining with LBP-TOP ☺

Weizmann downsampled videos – results



Spatial downsampling
$\mathrm{SD}_2, \mathrm{SD}_3$ *(k=2000)* & SD_4 *(k=1500)*

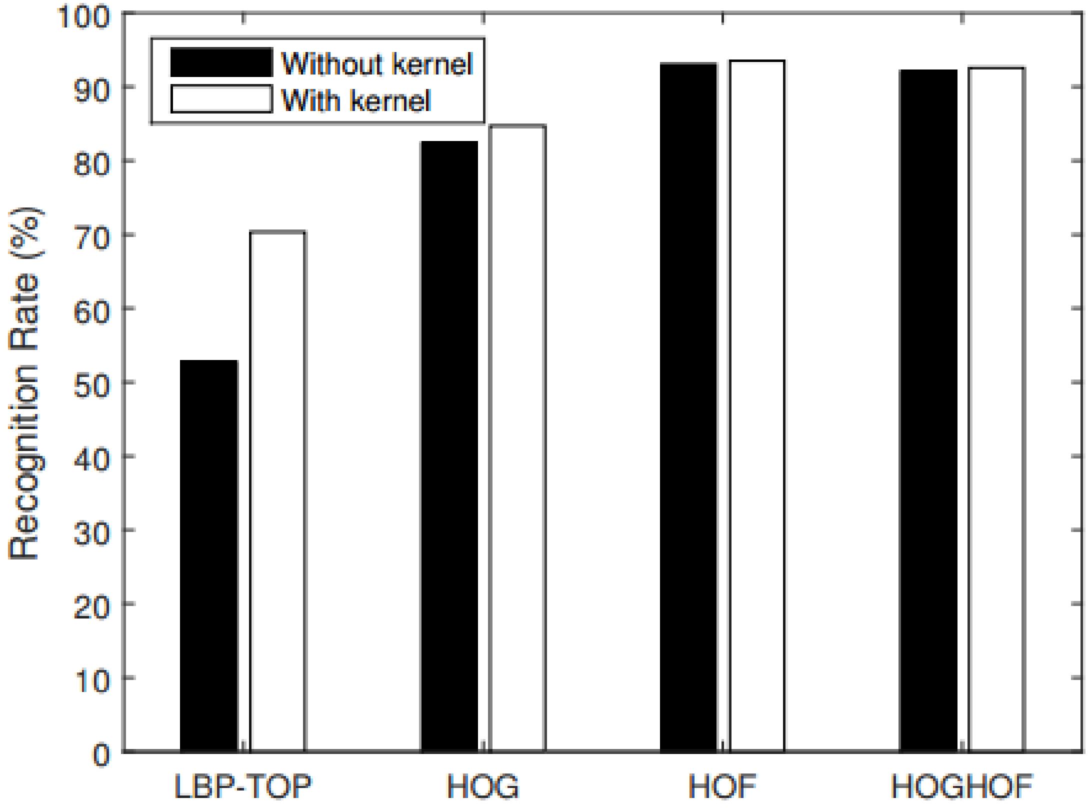


Temporal downsampling
SD_2 *(k=2000)*, SD_3 *(k=400)*

Weizmann downsampled videos – results (2)

- STIPs and kernalized LBP-TOP appear to dominate the best results within each mode ☺
- LBP-TOP contributes significantly more as the resolution quality decreases ☺
- Kernelized LBP-TOP achieves **best accuracy** rate at $\alpha = 4$ and $\beta = 3$ ☺

Effects of kernelization



Recognition accuracy with and without χ^2-kernel, on the original KTH videos.

Conclusion

- This work utilizes a new notion of joint feature utilization for action recognition in low quality videos
- This woks shows how downsampled videos can particularly get benefitted from textural information with shape and motion.
- The combined usage of all three features (HOG+HOF+LBP-TOP) outperforms the other competing methods across a majority of cases.
- Our best method is able to limit the drop in accuracy to around 8-10% when the video resolutions and frame rates deteriorate to a fourth of their original values.

Future Works

- Extend our evaluation to videos from more complex and uncontrolled environments [Laptev et al.'04], [Oh et al.'11]
- Investigate the simultaneous effects of both spatial and temporal downsampling on videos
- Explore other spatio-temporal textural features that might exhibit more robustness towards video quality

Thank You