Human Activity Recognition in Low Quality Videos using Spatio-Temporal Features

Saimunur Rahman

Masters (by Research) Viva

Thesis supervisor: Dr. John See Su Yang
Thesis co-supervisor: Dr. Ho Chiung Ching

Visual Processing Laboratory
Multimedia University, Cyberjaya

Introduction

Human Activity Recognition from Low Quality Videos

- Activity Recognition: Machine interpretation of human actions
  - Focus on low-level action primitives and actions of generic types
  - Examples: running, drinking, smoking, answering phone etc.

- Low Quality Video: Videos with poor quality settings
  - Low resolution and frame rate, camera motion, blurring, compression etc.




Video source: YouTube

Motivations & applications

- Existing frameworks does not assumes video quality as a problem
  - Designed for processing high quality videos
- Existing spatio-temporal representation methods are not robust to low quality videos
  - Not suitable for action modeling from lower quality videos
- Large application domains
  - Video search + indexing, surveillance applications,
  - Sports video analysis, dance choreography,
  - Human-computer interfaces, computer games etc.

Objectives of this research

Objective 1. To develop a framework for activity recognition in low quality videos

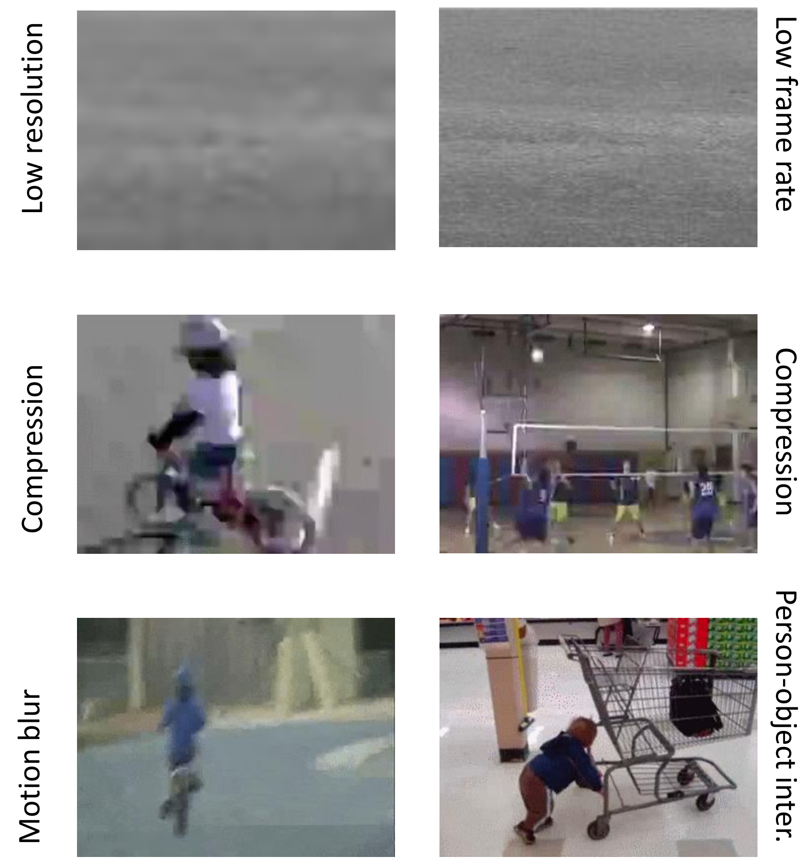
- Harness multiple spatio-temporal information in low quality videos
- Label a given video sequence as belonging to a particular action or not

Objective 2. To develop spatio-temporal feature representation method for activity recognition in low quality video

- Detect and encode spatio-temporal information inherit in videos
- Robust to low quality videos (much more challenging!)

Scope of Research

- Low quality videos
  - low spatial resolution
  - low sampling rate
  - compression artifacts
  - motion blur
- Type of human activities
  - single person activities
    - Ex. clapping, waving, running etc.
  - person-object interactions
    - Ex. hugging, playing basketball etc.



Video source: KTH actions [Schuld et al. 04], UCF-YouTube [Liu et al. 09], HMDB51 [Kuehne et al. 2011] and YouTube

Contributions of this research

- A framework for recognizing human activities in low quality videos
- A joint feature utilization method that combines shape, motion and textural features to improve the activity recognition performance
- A spatio-temporal mid level feature bank (STEM) for activity recognition in low quality videos
- Evaluations of recent shape, motion, and texture features and encoding methods on various low quality datasets.

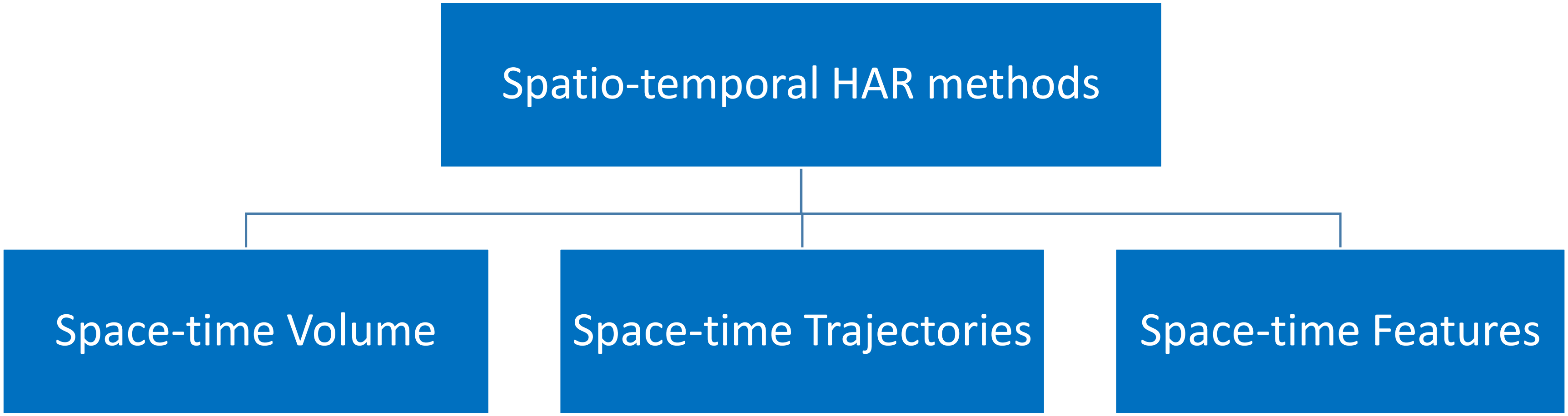
Presentation Outline

- Literature Review
- Dataset
- Joint Feature Utilization Method
- Spatio-temporal Mid-level Feature Bank
- Summary and Conclusion

Presentation Outline

- Literature Review
  - Thorough review of various state-of-the-art spatio-temporal feature representation methods
- Dataset
- Joint Feature Utilization Method
- Spatio-temporal Mid-level Feature Bank
- Summary and Conclusion

Literature Review



Space-time Volume (STV)

3D volume + template

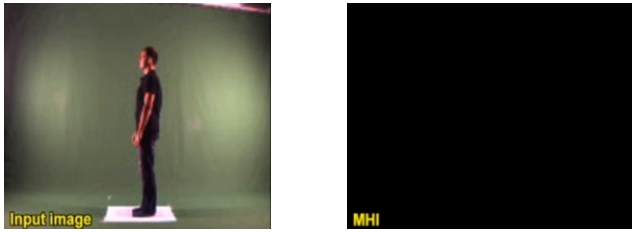
- MHI,MEI - Bobick and Davis (2001)
- GEI – Han & Bhanu (2006)
- MACH filter - Rodriguez et al. (2008)
- MHI + appearance – Hu et al. (2009)
- bMHI+ MHI contour - Qian et al. (2010)
- AMI - Kim et al. (2010)
- DMHI - Murakami (2010)
- GFI – Lam et al. (2011)
- Action Bank - Sadanand & Corso (2012)
- SFA – Zhang and Tao (2012)
- LPC- Shao and Tao (2014)
- LBP+MHI – Ahsan et al. (2014)
- OF+MHI - Tsai et al. (2015)
- EMF+GP – Shao et al. (2016)

Silhouette and skeleton

- HOR – Ikizler and Duygulu (2009)
- LPP – Fang et al. (2010)
- CSI – Ziaeefard & Ebrahimnezhad (2010)
- BB6-HM – Folgado et al. (2011)
- MHSV+TC – Karali & ElHelw (2012)
- BPH – Modarres & Soryani (2013)
- Action pose - Wang et al. (2013)
- Key pose - Chaaraoui (2013)
- Rep. & overw. MHI - Gupta et al. (2013)
- MoCap pose - Barnachon et al. (2014)
- STDE – Cheng et al. (2014)
- SPCI - Zhang et al. (2014)
- Shape+orient. - Vishwakarma et al (2015)
- MHI+TS - Lin et al. (2016)

Others

- CCA – Kim and Cipola (2009)
- HFM – Cao et al. (2009)
- PCA+SAU – Liu et al. (2010)
- 3D LSK – Seo & Milanfar (2011)
- DSA – Li et al. (2011)
- Grassmann manifolds - Harandi et al. (2013)
- PGA – Fu et al. (2013)
- Tensor decomposition - Su et al. (2014)
- CTW - Zhou & Torre (2016)



Input video source: Weizmann dataset, MHI [Bobick & Davis. (2001)]

- ❑ Use 3D (XYT) volume to model action
- ❑ Robust to noise and illumination changes
- ❑ Struggle to model activities with complex scenes
  - Not just simple periodic activities involving controlled environment
- ❑ Difficult to model activities if: resolution is low, multiple people interaction, over temporal downsampling

Space-time Trajectories (STT)

Salient Trajectories

- Harris3D+KLT - Messing et al. (2009)
- KLT tracker - Matikainen et al. (2009)
- SIFT matching - Sun et al. (2009)
- SIFT+KLT - Sun et al. (2010)
- ROI point - Raptis and Soatto (2010)
- Speech modeling - Chen & Aggarwal (2011)
- Weighted trajectories – Yu et al. (2014)

Dense Trajectories

- Dense traj. (DT) - Wang et al. (2011)
- DT+reference points – Jiang et al. (2012)
- Tracklet cluster trees – Gaidon et al. (2012)
- DT+FV - Atmosukarto et al. (2012)
- Improved DT (iDT) - Wang et al. (2013)
- DT+DCS – Jain et al. (2013)
- DT+context+mbh – Peng et al. (2013)
- iDT+SFV – Peng et al. (2013)
- Salient traj. – Yi & Lin (2013)
- TDD – Wang et al. (2015)
- Ordered traj. - Murthy & Goecke (2015)
- iDT+ img. CNN - Murthy & Goecke (2015)
- Web image CNN+iDT – Ma et al. (2016)

Others

- Chaotic invariants - Ali et al. (2007)
- Discriminative Topics Modelling - Bregonzio et al. (2010)
- Mid-Level action parts - Raptis et al. (2012)
- Harris3D+Graph - Aoun et al. (2014)
- local motion+group sparsity – Cho et al (2014)
- Dense body part - Murthy et al. (2014)

- ❑ Robust to the viewpoint and scale changes
- ❑ Computationally expensive
  - ❑ Tracking and feature matching is expensive
- ❑ Not suitable if spatial resolution is low or poor
  - ❑ Trajectories are estimated using spatial points

Input video source: YouTube

IDT [Wang et al. 13]

Space-time Features (STF)

STIPs

- Harris3D+Jet – Laptev (2005)
- Harris3D+Gradient – Laptev et al. (2008)
- Dollar+Cuboid – Dollar et al. (2008)
- Hessian+ESURF – Weilliams et al. (2008)
- Harris3D+HOG3D – Klaiser et al. (2009)
- Dollar+Gradient – Liu et al. (2009)
- Harris3D+LBP - Shao and Mattivi (2009)
- Harris3D+Gradeint - Kuehne et al. (2011)
- Feature mining - Gilbert et al. (2011)
- Action Bank – Sadanand & Corso (2012)
- Shape context - Zhao et al. (2013)
- Color STIP - Everts et al. (2014)
- Encoding Evaluations - Peng et al (2014)
- Harris3D+CNN - Murthy et al. (2015)

Dense Sampling

- Dense sampling (DS) – Wang et al. (2009)
- DS+HOG3D+SC – Zhu et al. (2010)
- Mid-level+DS - Liu et al (2012)
- Salient DS - Vig et al. (2013)
- Dense Tracklets – Bilinski et al. (2013)
- Saliency+DS - Vig et al. (2013)
- Real time strategy - Shi et al. (2013)
- DS+MBH - Peng et al. (2013)
- Real time DS - Uijlings et al. (2014)
- DS+HOG3D+LAG - Chen et al. (2015)
- STAP - Nguyen et al. (2015)
- DS+GBH - Shi et al. (2015)
- DS+LPM – Shi et al. (2016)

Unsupervisedly Learned

- CNN+LSTM – Baccouche et al. (2011)
- 3D CNN - Karpathy et al. (2014)
- Temporal Max Pooling - Ng et al. (2015)
- LRCN – Donahue et al. (2015)
- Two-stream CNN – Simonyan & Zisserman (2014)
- Multimodal CNN - Wu et al. (2015)
- Dynencoder – Yan et al. (2014)
- LSTM auto-encoder – Srivastava et al. (2015)
- Temporal coherence – Misra et al. (2016)
- Siamese Network – Wang et al. (2016)

- ❑ Suitable for modelling activities with complex scenes
- ❑ Robust to the scale changes
- ❑ Suitable for modeling multi-person interactions
- ❑ Struggles to handle viewpoint changes in the scenes
- ❑ Not suitable if image quality / structure is distorted


Input video


STIP [Laptev. 2003]

Video source: KTH dataset [Schuld et al. 2004]

Presentation Outline

- Literature Review
- Dataset
  - Overview and methodology for low quality version production
- Joint Feature Utilization Method
- Spatio-temporal Mid-level Feature Bank
- Summary and Conclusion

KTH Actions [Schüldt et al., 2004]

Dataset Description

- 6 action classes i.e. walking, running etc.
- Total 599 video samples
- Resolution: 160 $\times$ 120 pixels
- FPS: 25, Avg. clip: 10-15 sec.
- Evaluation: author specified test-train set.
- Result: average accuracy over all class

Spatial and temporal downsampling

SD_1	Original Res.	TD_1	Original F.R.
SD_2	Half res.	TD_2	Half F.R.
SD_3	One third res.	TD_3	One third F.R.
SD_4	One fourth res.	TD_4	One fourth F.R.

Spatially Downsampled

SD_1, SD_2, SD_3, SD_4

Temporally Downsampled

TD_1, TD_2, TD_3, TD_4

UCF-11 [Liu et al., 2009]

Dataset Description

- 11 action classes, 25 action groups
- Total 1600 videos
- Videos are affected by complex issues
- Resolution: 320 ╳ 240 pixels, 29.97 fps
- Evaluation: LOGOCV as per author
- Result: average accuracy over all class

Video Compression

- Re-encoded using x264 encoder
- Used CRF between 23 to 50 (referred as **YouTube-LQ**)
  - The higher the CRF the better compression !!
- Used uniform CRF[1] across all classes

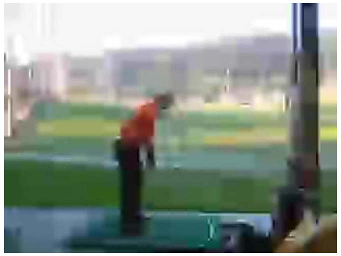
Original Videos

Compressed Videos

CRF 46

CRF 49

CRF 41



[1] The distribution of CRF values is available at http://saimunur.github.io/YouTube-LQ-CRFs.txt

HMDB51 [Kuehne et al., 2011]

Dataset Description

- 51 action classes
- Total 6766 videos
- Videos are affected by complex issues
- Quality metatag for video i.e. *good*, *medium*, *bad*
- Evaluation: test-training split by author
- Result: average accuracy over all class

Bad and Medium Quailty Videos

- Training with all videos in split
- Testing with only *'bad'* and *'medium'* quality videos i.e. **HMDB-BQ** and **HMDB-MQ**

HMDB-BQ

HMDB-MQ

Outline

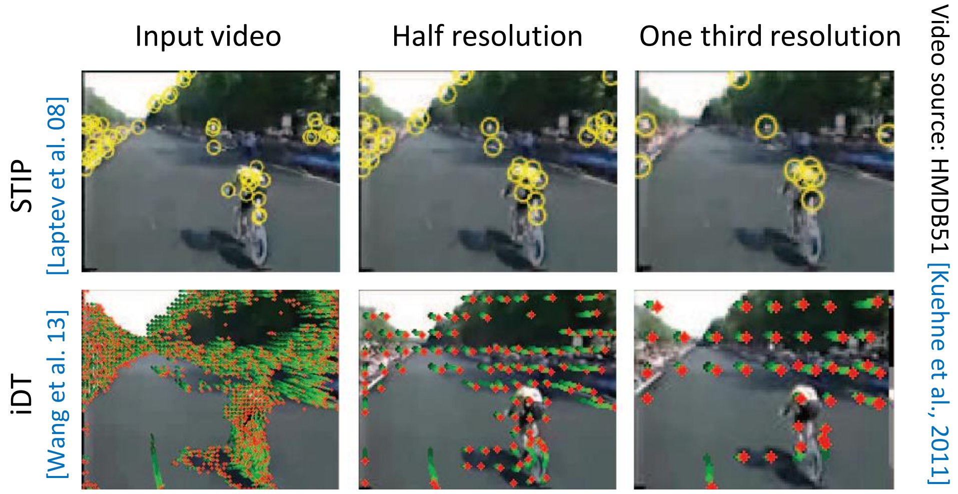
- Literature Review
- Dataset
- Joint Feature Utilization Method
  - Overview, motivation, feature representation methods, experimental results, and conclusion
- Spatio-temporal Mid-level Feature Bank
- Summary and Conclusion

Overview

- **Objective:** Joint feature utilization method for activity recognition in low quality videos
- **Main idea:** utilize shape, motion and textural features
  - Combine shape, motion and textures together
  - Alleviate individual shortcomings of each features for low quality videos
- **What is proposed?**
  - A feature fusion method of shape, motion and textural features
  - Textural features for improvement of state-of-the-art shape-motion features performance
  - A descriptor based on BSIF features [Kannala and Rahtu'11] for activity recognition

Motivation

- Shape-motion features does not perform well
  - Shape-motion feature detection is difficult if image is poor
  - Gradient changes (orientation+magnitute) are not significant enough
  - Global representation of statistical regularities is suitable in this situations



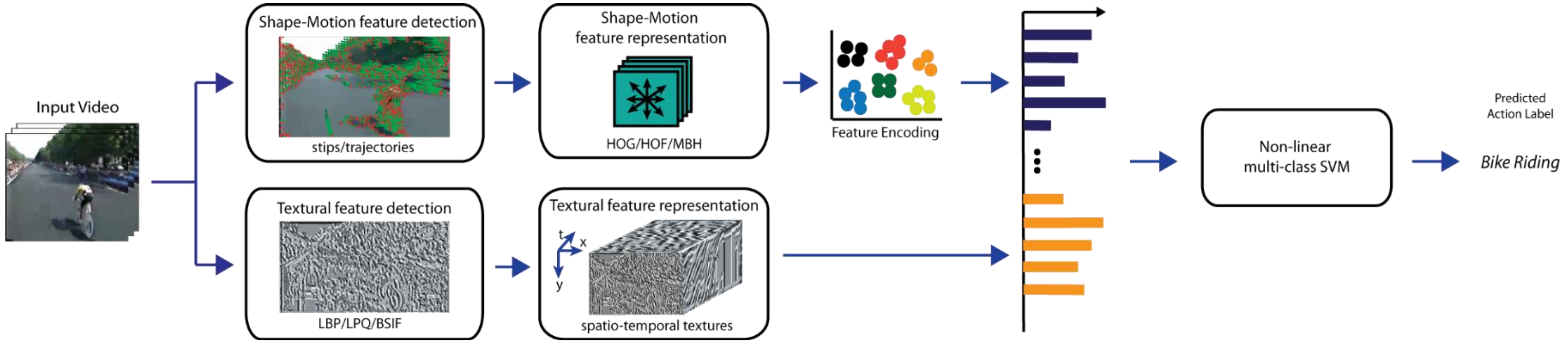
Performance of various detectors under different spatial quality condition

Spatio-temporal Feature Representation

- Shape and motion features
  - Space-time interest points (STIP) [Laptev et al. 08]
  - Improved dense trajectories (iDT) [Wang et al. 13, Wang et al. 15]
- Textural Features
  - Local Binary Pattern (LBP) [Ahonen et al. 06]
  - Local Phase Quantization (LPQ) [Ojansivu & Heikkilä. 08]
  - Binarized Statistical Image Features (BSIF) [Kannala and Rahtu. 11]
  - LBP, LPQ, BSIF are **lack of motion** (only captures shape information)
    - Three orthogonal plane (TOP) extension [Zhao et al. 08]

Joint Feature Utilization Framework

- Encode shape and motion features using BoVW model
- Encode textural features and concatenate with shape and motion feature histograms



Experimental results on KTH

Average accuracy (%)

Number of k-means clusters = 4000

VQ Encoding	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4
STIP (Baseline)	86.85	80.37	75.56	88.24	82.31	78.98
STIP + LBP-TOP	85.19	82.04	77.59	88.43	82.41	81.20
STIP + LPQ-TOP	87.41	80.19	76.30	87.41	81.85	79.81
STIP + BSIF-TOP	**88.80**	**85.28**	**81.67**	**88.70**	**86.11**	**84.54**

Number of GMM clusters = 256

FV Encoding	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4
STIP (Baseline)	89.44	82.41	79.07	88.89	85.74	85.09
STIP + LBP-TOP	**89.63**	82.69	78.52	**90.00**	85.65	**83.52**
STIP + LPQ-TOP	88.24	81.76	78.43	89.26	86.20	83.43
STIP + BSIF-TOP	89.26	**83.15**	**80.19**	89.91	**87.78**	82.96

Experimental results on KTH (2)

Average accuracy (%)

Number of k-means clusters = 4000

VQ Encoding	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4
iDT (Baseline)	92.59	78.80	61.85	95.19	91.57	89.54
iDT + LBP-TOP	92.96	81.94	73.61	95.09	92.13	89.54
iDT + LPQ-TOP	92.96	78.61	79.91	95.09	91.67	88.89
iDT + BSIF-TOP	**93.89**	**88.33**	**82.41**	95.09	**92.22**	**90.00**

Number of GMM clusters = 256

FV Encoding	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4
iDT (Baseline)	94.07	79.91	64.17	94.63	92.50	89.17
iDT + LBP-TOP	**94.26**	80.00	69.91	94.63	92.59	89.91
iDT + LPQ-TOP	94.07	80.00	78.80	94.63	92.59	89.63
iDT + BSIF-TOP	92.87	**87.78**	**81.02**	94.44	**92.59**	**90.28**

Average accuracy (%)

Experimental results on YouTube-LQ

Average accuracy (%)

Number of k-means clusters = 4000

Methods	VQ Encoding	FV Encoding
STIP (Baseline)	67.57	70.27
STIP + LBP-TOP	70.69	70.99
STIP + LPQ-TOP	69.13	71.65
STIP + BSIF-TOP	**76.05**	**75.04**

Average accuracy (%)

Methods	VQ Encoding	FV Encoding
iDT (Baseline)	74.04	67.10
iDT + LBP-TOP	75.59	68.57
iDT + LPQ-TOP	76.02	70.59
iDT + BSIF-TOP	**80.45**	**78.13**

Number of GMM clusters = 256

Average accuracy (%)

Experimental results on HMDB51

Average accuracy (%)

Number of k-means clusters = 4000

HMDB-BQ	VQ Encoding	FV Encoding
STIP (Baseline)	20.09	26.02
STIP + LBP-TOP	20.80	23.88
STIP + LPQ-TOP	23.89	25.02
STIP + BSIF-TOP	**32.46**	**33.06**

HMDB-MQ	VQ Encoding	FV Encoding
STIP (Baseline)	24.95	23.68
STIP + LBP-TOP	24.28	30.71
STIP + LPQ-TOP	28.36	30.75
STIP + BSIF-TOP	**37.14**	**38.51**

Number of GMM clusters = 256

HMDB-BQ	VQ Encoding	FV Encoding
iDT (Baseline)	28.87	30.98
iDT + LBP-TOP	30.34	30.57
iDT + LPQ-TOP	30.96	30.76
iDT + BSIF-TOP	**37.80**	**40.69**

HMDB-MQ	VQ Encoding	FV Encoding
iDT (Baseline)	41.43	46.35
iDT + LBP-TOP	43.11	45.43
iDT + LPQ-TOP	42.97	45.96
iDT + BSIF-TOP	**45.96**	**51.62**

Some Important Observations

- BSIF-TOP combinations (STIP+BSIF-TOP & iDT+BSIF-TOP) are superior then others
- Rank of texture performance: BSIF-TOP>LBP-TOP>LPQ-TOP
- iDT features and FV encoding performs better if quality of videos are good
- VQ encoding is better in case of spatially downsampled videos.

Conclusion

- A method for exploiting textural features into shape and motion features is proposed
- Use of textural features improves the recognition performance of shape-motion features by a good margin
  - Proposed BSIF-TOP performs better than other textures
- Evaluation of various feature combinations on various low quality datasets.
- Future work: more robust texture, rich texture feature description

Outline

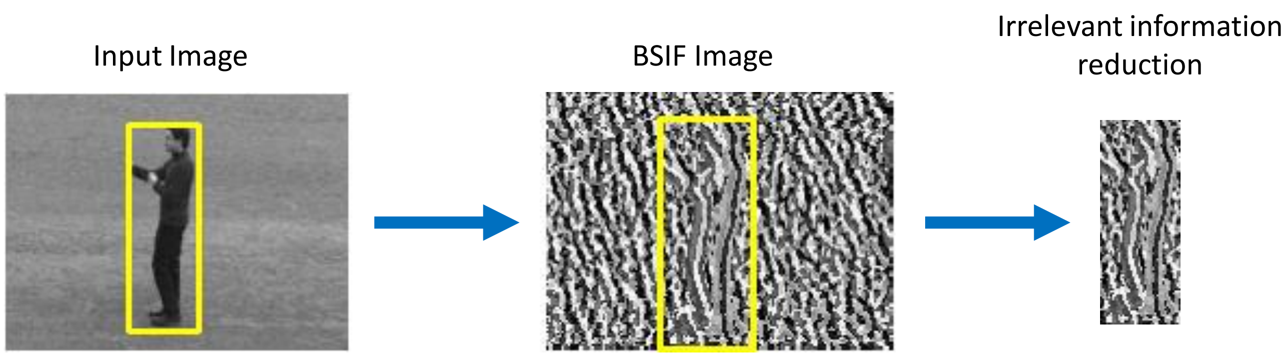
- Literature Review
- Dataset
- Joint Feature Utilization Method
- Spatio-temporal Mid-level Feature Bank
  - Overview, motivation, STEM overview, experimental results, and conclusion
- Summary and Conclusion

Overview

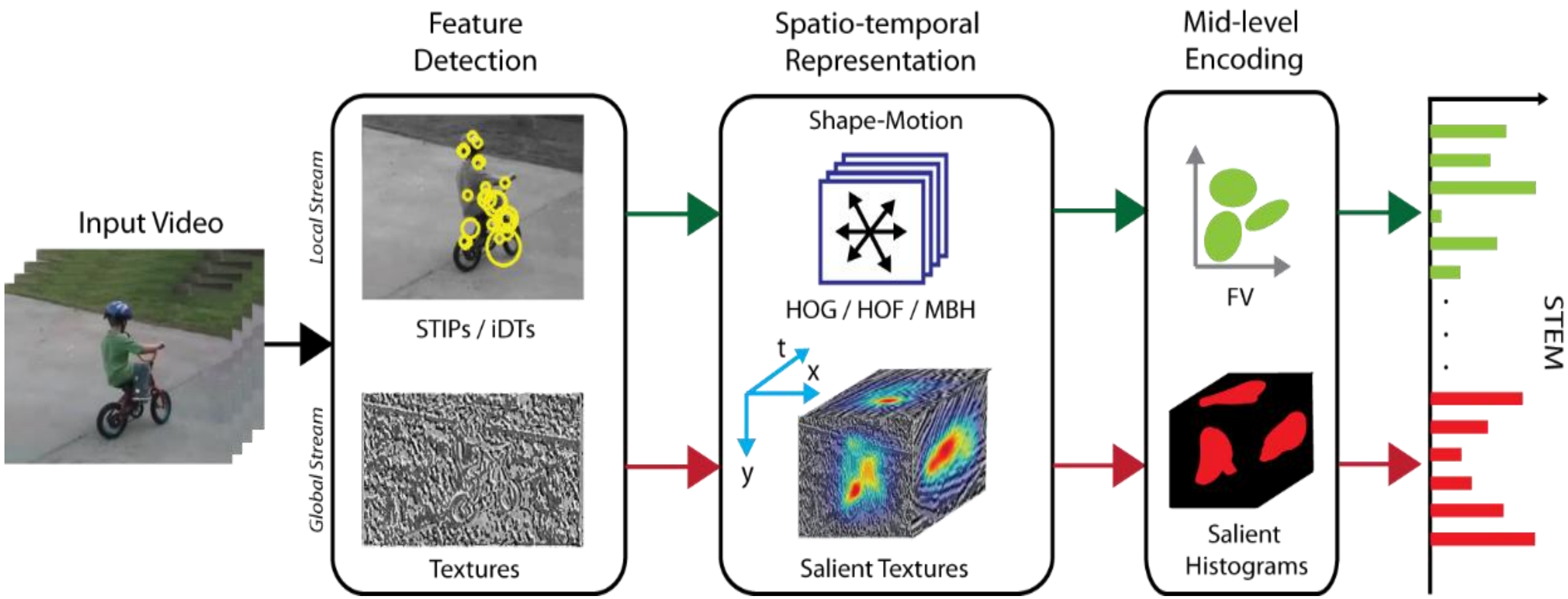
- **Objective:** a feature bank for low quality videos.
- **Main idea:** a feature bank consist of mid-level encoded features
  - Mid-level shape-motion features i.e. VQ vs. direct low-level features
  - Quantization of irrelevant textures reduce discriminative capacity of features
  - Intermediate pruning of textures (mid-level!!) removes irrelevant information
- **What is new?**
  - A new salient binarized image feature scheme
    - Used saliency map for removal of unnecessary features
  - Combine salient textures with shape-motion features

Motivation

- BSIF performs good with shape and motion features in low quality videos
- BSIF encodes many irrelevant and redundant information
- Reduction of irrelevant information increase the discriminative capacity



Spatio-temporal Mid-level Feature Bank



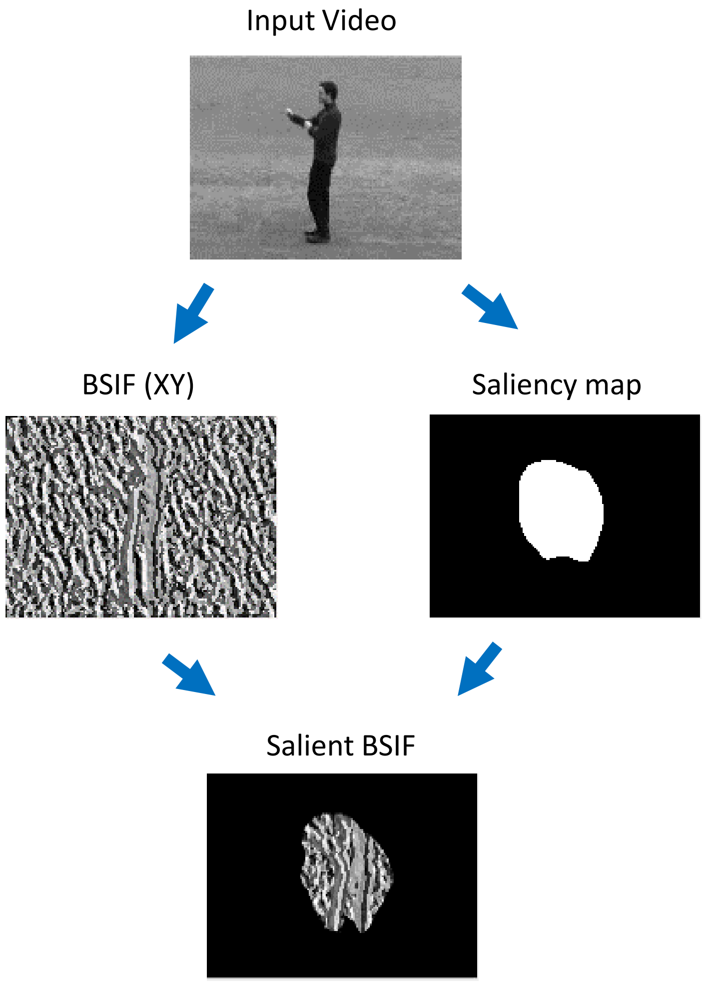
Spatio-temporal mid-level feature bank (STEM)

Shape-motion features

- Space-time interest point [Laptev 05]
  - Feature points are detected using Harris3D
  - A Cuboid is created around the feature point
  - Cuboid is described using gradient feature histogram (HOG and HOF)
- Feature trajectories [Wang et al. 13]
  - Trajectories are detected using Improved dense trajectories (original scale)
  - A trajectory aligned volume is created
  - Each volume is described using gradient feature histogram (MBH)

Salient textures

- Calculate BSIF image on XY, XT and YT plane
- Calculate corresponding saliency maps from input video using GBVS [Harel et al. 06]
- Convert saliency map to binary
  - Used Otsu method [Otsu. 75] for optimal threshold value
- Estimate salient BSIF features on XY, XT and YT plane
- Quantize BSIF features to form histogram



Experimental Results on KTH

Average accuracy (%)

Number of GMM clusters = 256

Methods	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4
STIP (Baseline)	89.44	82.41	79.07	88.89	85.74	85.09
STIP+LBP-TOP	89.63	82.69	78.52	90.00	85.65	83.52
$\text{STEM}_{\text{STIP}}$ (w/o saliency)	89.26	83.15	80.19	89.91	87.78	82.96
$\text{STEM}_{\text{STIP}}$	**90.28**	**83.61**	**82.96**	**89.81**	**88.24**	84.44

Methods	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4
iDT (Baseline)	94.07	79.91	64.17	94.63	92.50	89.17
iDT + LBP-TOP	94.26	80.00	69.91	94.63	92.59	89.91
STEM_{IDT} (w/o saliency)	92.87	87.78	81.02	94.44	92.59	90.28
STEM_{IDT}	**93.24**	**88.98**	**83.89**	94.54	**92.59**	89.81

Experimental Results on YouTube-LQ

Average accuracy (%)

Methods	Average Accuracy
STIP (Baseline)	70.27
STIP+LBP-TOP	70.99
$STEM_{STIP}$ (w/o saliency)	75.04
$STEM_{STIP}$	**77.49**

Methods	Average Accuracy
iDT (Baseline)	67.10
iDT+LBP-TOP	68.57
$STEM_{IDT}$ (w/o saliency)	78.13
$STEM_{IDT}$	**79.52**

Number of GMM clusters = 256

Experimental Results on HMDB51

Average accuracy (%)

Number of GMM clusters = 256

Methods	HMDB-BQ	HMDB-MQ
STIP (Baseline)	21.71	23.68
STIP+LBP-TOP	20.80	24.28
$STEM_{STIP}$ (w/o saliency)	32.46	37.14
$STEM_{STIP}$	31.69	**37.95**

Methods	HMDB-BQ	HMDB-MQ
iDT (Baseline)	30.98	46.35
iDT+LBP-TOP	30.57	45.43
$STEM_{IDT}$ (w/o saliency)	40.69	51.62
$STEM_{IDT}$	**40.92**	**51.79**

Comparison with state-of-art

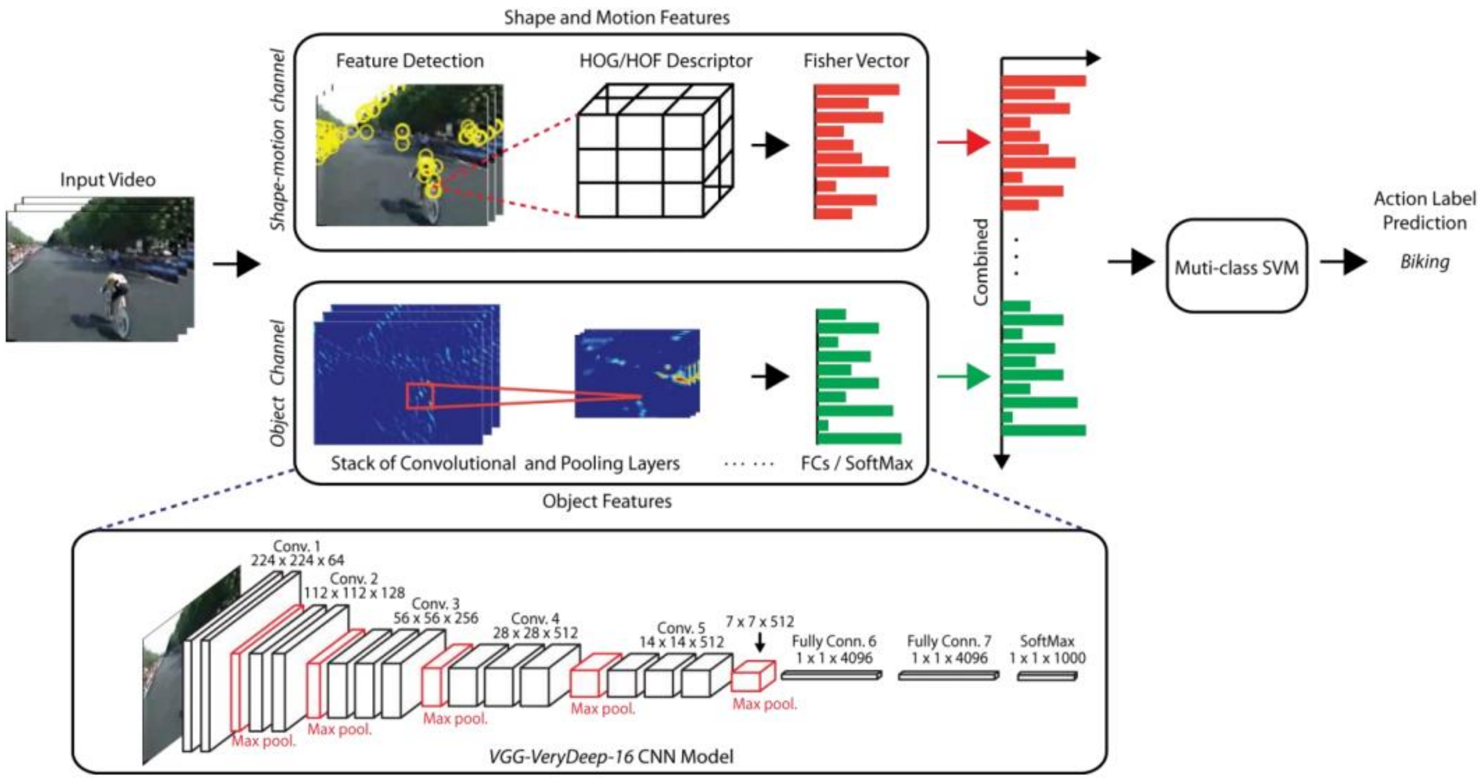
Average accuracy (%)

Methods	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4	YouTube-LQ	HMDB-BQ	HMDB-MQ
STIP [Wang et al. 09]	87.96	79.63	75.00	85.19	79.17	77.31	63.88	17.04	22.77
HOG+HOF [Wang et al. 13]	89.44	82.41	79.07	88.89	85.74	85.09	70.27	21.71	23.68
iDT(MBH) [Wang et al. 13]	92.59	78.80	61.85	95.19	91.57	89.54	67.10	30.98	46.35
STIP+LBP-TOP [See & Rahman 15]	89.81	81.48	78.70	89.35	86.11	84.72	70.99	20.80	24.28
$STEM_{STIP}$ (w/o saliency)	89.26	83.15	80.19	89.91	87.78	82.96	75.04	32.46	37.14
$STEM_{IDT}$ (w/o saliency)	92.87	87.78	81.02	94.44	92.59	90.28	78.13	40.69	51.62
$STEM_{STIP}$	90.28	83.61	**82.96**	89.81	88.24	84.44	**77.49**	31.69	**37.95**
$STEM_{IDT}$	**93.24**	**88.98**	**83.89**	**94.54**	**92.59**	89.81	**79.52**	**40.92**	**51.79**

Conclusion

- A spatio-temporal mid-level feature bank (STEM) was proposed
  - Integrate advantage of local interest points and global salient patches
- Proposed method performed well in various low quality datasets
- STEM can be further improved by multi-scale BSIF-TOP expansion
- Future work: robust saliency method, prune shape-motion features

Additional Experiments (1)



Deep Object Features for Improved Action Recognition

- **Shape-motion Channel:** Harris3D + HOG/HOF
- **Object Channel:** VGG-16 trained on ImageNet + FCs/SoftMax
- **Classification:** multi-class SVM + chi^2 homogeneous kernel

Additional Experiments (2)

Average accuracy (%)

Method	YouTube-LQ	HMDB51-BQ	HMDB51-MQ
HOG+HOF+LBP-TOP	70.99	23.88	30.71
HOG+HOF+LPQ-TOP	71.65	25.02	30.75
STEM (w/o saliency)	75.04	33.78	38.76
STEM	77.50	**34.08**	38.94
HOG+FC6+FC7	**84.03**	33.02	**40.05**
HOF+FC6+FC7	**85.16**	32.80	**40.41**
HOG+HOF+FC6+FC7	**86.34**	33.74	**40.55**

Shape, Motion and Object Features Vs. STEM and JFU

Method	YouTube-LQ	HMDB51-BQ	HMDB51-MQ
Softmax	77.42	23.31	30.46
FC6	**83.54**	23.31	30.50
FC7	**81.33**	28.41	38.02
FC6+FC7	**83.13**	31.99	**39.63**
FC6+FC7+softmax	**83.08**	31.98	**39.70**

Individual and combination of Deep Object Features

Outline

- Literature Review
- Dataset
- Joint Feature Utilization Method
- Spatio-temporal Mid-level Feature Bank
- Summary and Conclusion

Summary

- Several contributions to activity recognition (framework and feature representation) in low quality video settings have been presented
  - A framework for feature extraction and representation
  - A joint feature utilization method that involves utilization of shape-motion and textural features
  - A spatio-temporal mid-level feature bank that discriminately extracts salient textural features
  - Evaluation of state-of-the-art methods for low quality video

Future Work

- Joint Feature Utilization
  - Design features specific to poor quality
  - Further exploration of BSIF like features
- Mid-level feature bank
  - Saliency map robust to complex scenes
    - Deep learning for saliency map
  - Pruning shape-motion features using saliency map
- Unsupervised feature representation
  - CNN features recently showed good results for video classification

Publications (International Conference)

1. Saimunur Rahman, John See and Chiung Ching Ho (2016b). Deep Object features for improved action recognition in low quality videos. In *International conference on Computational Science and Engineering* (ICCSE), pp. To appear. [ISI-Scopus indexed].

2. Saimunur Rahman and John See (2016a). Spatio-temporal mid-level feature bank for action recognition in low quality video. In IEEE *International conference on Acoustics, speech and signal processing* (ICASSP 2016), pp. To appear. [CORE B].

3. Saimunur Rahman, John See and Chiung Ching Ho (2016a). Leveraging textural features for recognizing actions in low quality videos. In *International conference on Robotics, vision, signal processing and power applications* (ROVISP), pp. To appear. [ISI-Scopus indexed].

4. John See & Saimunur Rahman (2015b). On the effects of low video quality in human action recognition. In *international conference on Digital image computing: Techniques and applications* (DICTA), pp. 1-8. [CORE B].

5. Saimunur Rahman, John See and Chiung Ching Ho (2015b). Action recognition in low quality videos by jointly using shape, motion and texture features. In *international conference on Signal and image processing applications* (ICSIPA), pp. 83–88. [ISI-Scopus indexed].

Publications (Under Review)

1. Saimunur Rahman, John See, & Chiung Ching Ho. (2016). Joint feature utilization for human action recognition in low quality videos. *Journal of Neurocomputing* (SJR Q2).

2. Saimunur Rahman, John See, & Chiung Ching Ho. (2016). A review on spatio-temporal features for human action recognition. *International Journal of Pattern Recognition and Artificial Intelligence* **(IJPRAI)** (SJR Q2).

3. Saimunur Rahman and John See. (2016). A three-stream network for human action recognition in low quality videos. *Journal of Image and Vision Computing* (SJR Q1).

Acknowledgements

- All my friends and colleagues at Multimedia University
- All members of Centre of Visual Computing
- All of panel members during my proposal and work completion denfense
- Ministry of Higher Education for FRGS research grant
- Multimedia University for top-up scholarship

Huge thanks to

Dr. John See

Dr. Peter Ho

For introducing me to the visual computing world, its amazing!

References (1)

1. Aggarwal, J. K., & Cai, Q. (1997). Human motion analysis: A review. In *Nonrigid and articulated motion workshop, 1997. proceedings., ieee* (pp. 90–102).
2. Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, *43*(3), 16.
3. Ahad, M. A., Tan, J., Kim, H., & Ishikawa, S. (2010). A simple approach for low-resolution activity recognition. *Int. J. Comput. Vis. Biomech*, *3*(1).
4. Ahad, M. A. R., Ogata, T., Tan, J., Kim, H., & Ishikawa, S. (2008). A complex motion recognition technique employing directional motion templates. *International Journal of Innovative Computing, Information and Control*, *4*(8), 1943–1954.
5. Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *28*(12), 2037–2041.
6. Ahsan, S. M. M., Tan, J. K., Kim, H., & Ishikawa, S. (2014). Histogram of dmhi and lbp images to represent human actions. In *Image processing (icip), 2014 ieee international conference on* (pp. 1440–1444).
7. Baumann, F., Ehlers, A., Rosenhahn, B., & Liao, J. (2016). Recognizing human actions using novel space-time volume binary patterns. *Neurocomputing*, *173*, 54–63.
8. Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as space-time shapes. In *Computer vision, 2005. iccv 2005. tenth ieee international conference on* (Vol. 2, pp. 1395–1402).
9. Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in neural information processing systems*, *18*, 147.
10. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.
11. Bobick, A., & Davis, J. (1996). An appearance-based representation of action. In *Pattern recognition, 1996., proceedings of the 13th international conference on* (Vol. 1, pp. 307–312).
12. Bobick, A. F. (1997). Movement, activity and action: the role of knowledge in the perception of motion. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *352*(1358), 1257–1265.
13. Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *23*(3), 257–267.
14. Boureau, Y.-L., Ponce, J., & LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (icml-10)* (pp. 111–118).
15. Bregonzio, M., Gong, S., & Xiang, T. (2009). Recognising action as clouds of space-time interest points. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp.1948–1955).

References (2)

16. Cao, L., Luo, J., Liang, F., & Huang, T. S. (2009). Heterogeneous feature machines for visual recognition. In *Computer vision, 2009 ieee 12th international conference on* (pp. 1095–1102).
17. Chakraborty, B., Holte, M. B., Moeslund, T. B., & Gonzàlez, J. (2012). Selective spatio-temporal interest points. *Computer Vision and Image Understanding*, *116*(3), 396–410.
18. Chen, C.-C., & Aggarwal, J. (2009). Recognizing human action from a far field of view. In *Motion and video computing, 2009. wmvc'09. workshop on* (pp. 1–7).
19. Chen, C.-C., & Aggarwal, J. (2011). Modeling human activities as speech. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on* (pp. 3425–3432).
20. Chen, X., Cheng, Y., & Yi, Y. (2015). Features extraction approach based on dense salient trajectories in videos. In *Bioelectronics and bioinformatics (isbb), 2015 international symposium on* (pp. 132–135).
21. Cheng, G., Wan, Y., Saudagar, A. N., Namuduri, K., & Buckles, B. P. (2015). Advances in human action recognition: A survey. *arXiv preprint arXiv:1501.05964*.
22. Dawn, D. D., & Shaikh, S. H. (2015). A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector. *The Visual Computer*, 1–18.
23. Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatiotemporal features. In *Visual surveillance and performance evaluation of tracking and surveillance, 2005. 2nd joint ieee international workshop on* (pp. 65–72).
24. Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). Recognizing action at a distance. In *Computer vision, 2003. proceedings. ninth ieee international conference on* (pp. 726–733).
25. Fang, C.-H., Chen, J.-C., Tseng, C.-C., & Lien, J.-J. J. (2009). Human action recognition using spatio-temporal classification. In *Computer vision–accv 2009* (pp. 98–109). Springer.
26. Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *32*(9), 1627–1645.
27. Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395.
28. Gavrila, D. M. (1999). The visual analysis of human movement: A survey. *Computer vision and image understanding*, *73*(1), 82–98.
29. Guo, K., Ishwar, P., & Konrad, J. (2010). Action change detection in video by covariance matching of silhouette tunnels. In *Acoustics speech and signal processing (icassp), 2010 ieee international conference on* (pp. 1110–1113).
30. Han, J., & Bhanu, B. (2006). Individual recognition using gait energy image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *28*(2), 316–322.

References (3)

31. Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Advances in neural information processing systems* (pp. 545–552).
32. Harjanto, F., Wang, Z., Lu, S., Tsoi, A. C., & Feng, D. D. (2015). Investigating the impact of frame rate towards robust human action recognition. *Signal Processing*.
33. Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Alvey vision conference* (Vol. 15, p. 50).
34. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., & Huang, T. S. (2009). Action detection in complex scenes with spatial and temporal ambiguities. In *Computer vision, 2009 ieee 12th international conference on* (pp. 128–135).
35. Ikizler, N., Cinbis, R. G., & Duygulu, P. (2008). Human action recognition with line and flow histograms. In *Pattern recognition, 2008. icpr 2008. 19th international conference on* (pp. 1–4).
36. Ikizler, N., & Duygulu, P. (2009). Histogram of oriented rectangles: A new pose descriptor for human action recognition. *Image and Vision Computing*, *27*(10), 1515–1526.
37. Jaakkola, T. S., Haussler, D., et al. (1999). Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, 487–493.
38. Jain, A., Gupta, A., Rodriguez, M., & Davis, L. (2013). Representing videos using mid-level discriminative patches. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2571–2578).
39. Johansson, G. (1975). Visual motion perception. *Scientific American*.
40. Kannala, J., & Rahtu, E. (2012). Bsif: Binarized statistical image features. In *Pattern recognition (icpr), 2012 21st international conference on* (pp. 1363–1366).
41. Kataoka, H., Aoki, Y., Iwata, K., & Satoh, Y. (2015a). Evaluation of vision-based human activity recognition in dense trajectory framework. In *Advances in visual computing* (pp. 634–646). Springer.
42. Kataoka, H., Aoki, Y., Iwata, K., & Satoh, Y. (2015b). Evaluation of vision-based human activity recognition in dense trajectory framework. In *Visual computing (isvc), 11th international symposium on* (p. To Appear).
43. Ke, S.-R., Thuc, H. L. U., Lee, Y.-J., Hwang, J.-N., Yoo, J.-H., & Choi, K.-H. (2013). A review on video-based human activity recognition. *Computers*, *2*(2), 88–131.
44. Kellokumpu, V., Zhao, G., & Pietikäinen, M. (2008). Human activity recognition using a dynamic texture based method. In *Bmvc* (Vol. 1, p. 2).
45. Kellokumpu, V., Zhao, G., & Pietikäinen, M. (2011). Recognition of human actions using texture descriptors. *Machine Vision and Applications*, *22*(5), 767–780.

References (4)

46. Kim, T.-K., & Cipolla, R. (2009). Canonical correlation analysis of video volume tensors for action categorization and detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*(8), 1415–1428.
47. Kim, W., Lee, J., Kim, M., Oh, D., & Kim, C. (2010). Human action recognition using ordinal measure of accumulated motion. *EURASIP journal on Advances in Signal Processing*, *2010*(1), 1–11.
48. Klaser, A., Marszałek, M., & Schmid, C. (2008). A spatio-temporal descriptor based on 3dgradients. In *Bmvc 2008-19th british machine vision conference* (pp. 275–1).
49. Koenderink, J. J., & van Doorn, A. J. (1987). Representation of local geometry in the visual system. *Biological cybernetics*, *55*(6), 367–375.
50. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). Hmdb: a large video database for human motion recognition. In *Computer vision (iccv), 2011 ieee international conference on* (pp. 2556–2563).
51. Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, *64*(2-3), 107–123.
52. Laptev, I., & Lindeberg, T. (2003). Space-time interest points. In *In iccv* (pp. 432–439).
53. Laptev, I., Marszałek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Computer vision and pattern recognition, 2008. cvpr 2008. ieee conference on* (pp. 1–8).
54. Lin, Z., Jiang, Z., & Davis, L. S. (2009). Recognizing actions by shape-motion prototype trees. In *Computer vision, 2009 ieee 12th international conference on* (pp. 444–451).
55. Lindeberg, T. (1998). Feature detection with automatic scale selection. *International journal of computer vision*, *30*(2), 79–116.
56. Liu, C., & Yuen, P. C. (2010). Human action recognition using boosted eigenactions. *Image and vision computing*, *28*(5), 825–835.
57. Liu, J., Luo, J., & Shah, M. (2009). Recognizing realistic actions from videos “in the wild”. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 1996–2003).
58. Lu, W.-L., & Little, J. J. (2006). Simultaneous tracking and action recognition using the pca-hog descriptor. In *Computer and robot vision, 2006. the 3rd canadian conference on* (pp. 6–6).
59. Lucas, B. D., Kanade, T., et al. (1981). An iterative image registration technique with an application to stereo vision. In *Ijcai* (Vol. 81, pp. 674–679).
60. Mattivi, R., & Shao, L. (2009). Human action recognition using lbp-top as sparse spatio-temporal feature descriptor. In *Computer analysis of images and patterns* (pp. 740–747).

References (5)

61. Messing, R., Pal, C., & Kautz, H. (2009). Activity recognition using the velocity histories of tracked keypoints. In *Computer vision, 2009 ieee 12th international conference on* (pp. 104–111).
62. Mitra, P., Murthy, C., & Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *24*(3), 301–312.
63. Moeslund, T. B., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, *104*(2), 90–126.
64. Murthy, O., & Goecke, R. (2013). Ordered trajectories for large scale human action recognition. In *Proceedings of the ieee international conference on computer vision workshops* (pp. 412–419).
65. Murthy, O. R., & Goecke, R. (2015). Ordered trajectories for human action recognition with large number of classes. *Image and Vision Computing*, *42*, 22–34.
66. Nanni, L., Lumini, A., & Brahnam, S. (2012). Survey on lbp based texture descriptors for image classification. *Expert Systems with Applications*, *39*(3), 3634–3641.
67. Ojansivu, V., & Heikkilä, J. (2008). Blur insensitive texture classification using local phase quantization. In *Image and signal processing* (pp. 236–243). Springer.
68. Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, *11*(285- 296), 23–27.
69. Päivärinta, J., Rahtu, E., & Heikkilä, J. (2011). Volume local phase quantization for blurinsensitive dynamic texture classification. In *Image analysis* (pp. 360–369). Springer.
70. Peng, X., Wang, L., Wang, X., & Qiao, Y. (2014). Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *arXiv preprint arXiv:1405.4506*.
71. Perronnin, F., Sánchez, J., & Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *Computer vision–eccv 2010* (pp. 143–156). Springer.
72. Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, *28*(6), 976–990.
73. Qian, H., Mao, Y., Xiang, W., & Wang, Z. (2010). Recognition of human activities using svm multi-class classifier. *Pattern Recognition Letters*, *31*(2), 100–111.
74. Reddy, K. K., Cuntoor, N., Perera, A., & Hoogs, A. (2012). Human action recognition in largescale datasets using histogram of spatiotemporal gradients. In *Advanced video and signal-based surveillance (avss), 2012 ieee ninth international conference on* (pp. 106–111).

References (6)

75. Roh, M.-C., Shin, H.-K., & Lee, S.-W. (2010). View-independent human action recognition with volume motion template on single stereo camera. *Pattern Recognition Letters*, *31*(7), 639–647.
76. Ryoo, M., Chen, C.-C., Aggarwal, J., & Roy-Chowdhury, A. (2010). An overview of contest on semantic description of human activities (sdha) 2010. In *Recognizing patterns in signals, speech, images and videos* (pp. 270–285). Springer.
77. Sadanand, S., & Corso, J. J. (2012). Action bank: A high-level representation of activity in video. In *Computer vision and pattern recognition (cvpr), 2012 ieee conference on* (pp. 1234–1241).
78. Schüldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local svm approach. In *Pattern recognition, 2004. icpr 2004. proceedings of the 17th international conference on* (Vol. 3, pp. 32–36).
79. See, J., & Rahman, S. (2015). On the effects of low video quality in human action recognition. In *Digital image computing: Techniques and applications (dicta), 2015 international conference on* (pp. 1–8).
80. Sultani, W., & Saleemi, I. (2014). Human action recognition across datasets by foregroundweighted histogram decomposition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 764–771).
81. Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, *18*(11), 1473–1488.
82. Vedaldi, A., & Zisserman, A. (2012). Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *34*(3), 480–492.
83. Vili, K., Guoying, Z., & Matti, P. (2008). Texture based description of movements for activity analysis. In *Int. conf. on computer vision theory and applications (visapp 2008)* (Vol. 1, pp. 206–213).
84. Vishwakarma, S., & Agrawal, A. (2013). A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, *29*(10), 983–1009.
85. Wang, H., Kläser, A., Schmid, C., & Liu, C.-L. (2011). Action recognition by dense trajectories. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on* (pp. 3169–3176).
86. Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In *Proceedings of the ieee international conference on computer vision* (pp. 3551–3558).
87. Wang, H., Ullah, M. M., Klaser, A., Laptev, I., & Schmid, C. (2009). Evaluation of local spatiotemporal features for action recognition. In *Bmvc 2009-british machine vision conference* (pp. 124–1).
88. Wang, L., Hu, W., & Tan, T. (2003). Recent developments in human motion analysis. *Pattern recognition*, *36*(3), 585–601.

References (6)

89. Wang, L., Qiao, Y., & Tang, X. (2013). Motionlets: Mid-level 3d parts for human motion recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2674–2681).
90. Wang, L., Qiao, Y., & Tang, X. (2014). Action recognition and detection by combining motion and appearance features. *THUMOS14 Action Recognition Challenge*, *1*, 2.
91. Wang, X., Wang, L., & Qiao, Y. (2013). A comparative study of encoding, pooling and normalization methods for action recognition. In *Computer vision–accv 2012* (pp. 572–585). Springer.
92. Wang, Y., & Mori, G. (2009). Human action recognition by semilatent topic models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*(10), 1762–1774.
93. Wiegand, T., Sullivan, G. J., Bjøntegaard, G., & Luthra, A. (2003). Overview of the h. 264/avc video coding standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, *13*(7), 560–576.
94. Willems, G., Tuytelaars, T., & Van Gool, L. (2008). An efficient dense and scale-invariant spatiotemporal interest point detector. In *Computer vision–eccv 2008* (pp. 650–663). Springer.
95. Wu, Q., Wang, Z., Deng, F., Xia, Y., Kang, W., & Feng, D. D. (2013). Discriminative two-level feature selection for realistic human action recognition. *Journal of Visual Communication and Image Representation*, *24*(7), 1064–1074.
96. Wu, X., Xu, D., Duan, L., & Luo, J. (2011). Action recognition using context and appearance distribution features. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on* (pp. 489–496).
97. Wu, X., Xu, D., Duan, L., Luo, J., & Jia, Y. (2013). Action recognition using multilevel features and latent structural svm. *Circuits and Systems for Video Technology, IEEE Transactions on*, *23*(8), 1422–1431.
98. Xu, H., Tian, Q., Wang, Z., & Wu, J. (2015). A survey on aggregating methods for action recognition with dense trajectories. *Multimedia Tools and Applications*, 1–17.
99. Xu, X., Tang, J., Zhang, X., Liu, X., Zhang, H., & Qiu, Y. (2013). Exploring techniques for vision based human activity recognition: Methods, systems, and evaluation. *Sensors*, *13*(2), 1635–1650.
100. Yeffet, L., & Wolf, L. (2009). Local trinary patterns for human action recognition. In *Computer vision, 2009 ieee 12th international conference on* (pp. 492–497).
101. Yi, Y., & Lin, Y. (2013). Human action recognition with salient trajectories. *Signal processing*, *93*(11), 2932–2941.
102. Yuan, C., Li, X., Hu, W., Ling, H., & Maybank, S. (2013, June). 3d r transform on spatiotemporal interest points for action recognition. In *The ieee conference on computer vision and pattern recognition (cvpr).*
103. Zhang, D., & Zhou, Z.-H. (2005). (2d) 2pca: Two-directional two-dimensional pca for efficient face representation and recognition. *Neurocomputing*, *69*(1), 224–231.
104. Zhao, D., Shao, L., Zhen, X., & Liu, Y. (2013). Combining appearance and structural features for human action recognition. *Neurocomputing*, *113*, 88–96.
105. Zhao, G., & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *29*(6), 915–928.
106. Ziaeefard, M., & Ebrahimnezhad, H. (2010). Hierarchical human action recognition by normalized-polar histogram. In *Pattern recognition (icpr), 2010 20th international conference on* (pp. 3720–3723).

Thank you for your attention

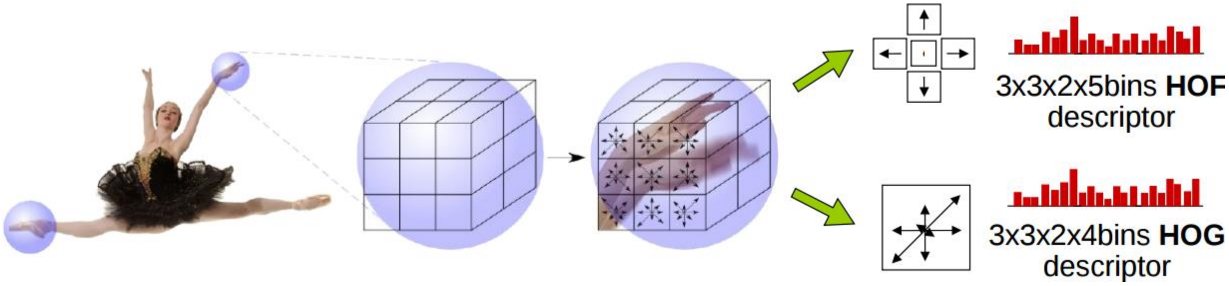
Any Questions?

Space-time interest points (STIP)

[Laptev et al. 08]

- Interest point (IP) detection
  - Harris3D detector
- Feature description
  - A cuboid of around interest point is created
  - Cuboid is divided into $n_x \times n_y \times n_t$ cells
  - Each cell is described using HOG (4-bins) and HOF (5-bins)
  - The size of descriptor: $\Delta_x(\sigma) = \Delta_y(\sigma) = 18\sigma, \Delta_t(\tau) = 8\tau$
    - σ is spatial scale and τ is temporal scale i.e. $\sigma = 3, \tau = 2$

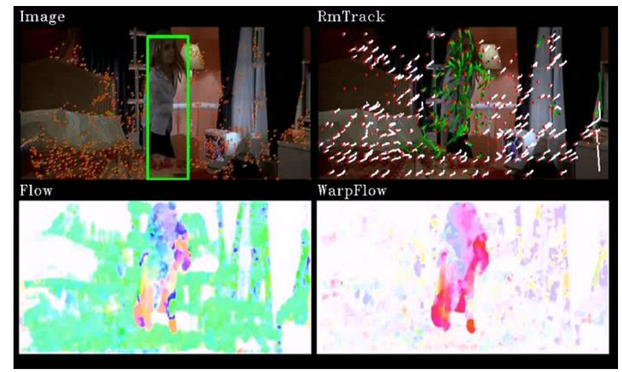
Harris3D in action



Improved dense trajectories (iDT) [Wang et al. 13]

- Camera motion removal
  - Homography estimation using RANSAC [Fischler & Bolles. 1981]
    - SURF and Optical flow (OF) for similarity between two frames
  - Re-compute the optical flow – *warped flow*
- Trajectory estimation
  - Trajectories using dense trajectories [Wang et al. 11]
  - Track points with original spatial scale (2-3% less than multi-scale)
- Trajectory aligned feature description
  - A cuboid of N cells across the trajectory length *L*×*L*
  - Cuboid is divided into $n_x \times n_y \times n_t$ cells.
  - For each cell a 8-bin histogram for both MBHx and MBHy
  - Size of descriptor: $\Delta_x(\sigma) = \Delta_y(\sigma) = 32\sigma, \Delta_t(\tau) = 3\tau$
    - σ is spatial scale and τ is temporal scale i.e. $\sigma = 2, \tau = 3$

Input video **Trajectory detection**



OF with motion **OF without motion**

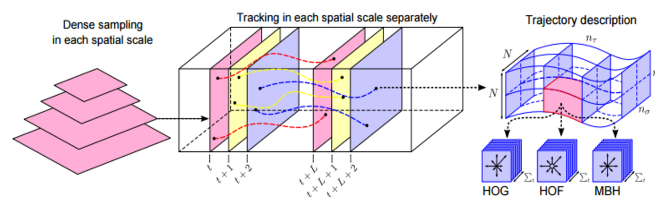


Image reproduced from Wang et al. 2011

Input video source: YouTube

Local Binary Pattern

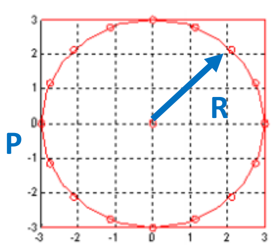
[Ahonen et al. 06]

- Describe each image pixel by relative grey levels of its neighbourhood pixels

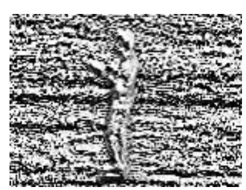
$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \qquad s(x) = \begin{cases} 1, if\ x \ \geq \ 0; \\ 0, otherwise. \end{cases}$$

g_c = graylevel of the centre pixel
g_p = N equally spaced neighbourhood pixel

- Produces 2^P different binary pattern
- The final feature histogram is from LBP output values



Input image



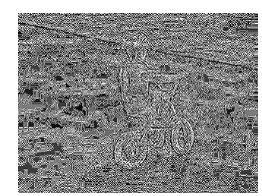
LBP image

Local Phase Quantization

[Ojansivu & Heikkilä 08]

- Use short term Fourier transform (STFT) in rectangular neighbourhood N_x
- Four complex coefficients are calculated
- 8 binary coefficient is formed form the sign of imaginary and real part
- An image representing 8 binary values is formed
- The final feature histogram is from LPQ image

Input image



LPQ image

Binarized statistical image features (BSIF)

[Kannala & Rahtu 2012]

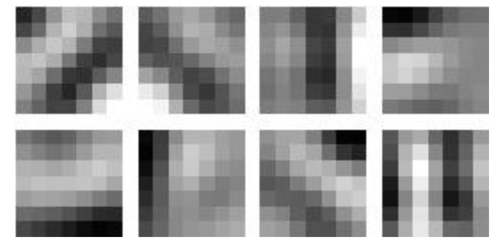
- Use linear filter F_i learnt from natural images through independent component analysis (ICA)

$$r_i = \sum_{u,v} F_i(u,v)X(u,v) = \mathbf{f}_i^T \mathbf{x}$$

- Binarized features b_i:

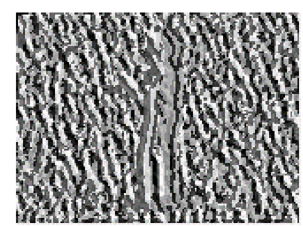
$$b_i = \begin{cases} 1; \ r_i > 0 \\ 0; otherwise \end{cases}$$

- n-bit binary code is produced for each pixel and form an image
- The final feature histogram is from BSIF image



9x9 8-bit learned filter

Input video



Output BSIF video using 9x9 8-bit filter

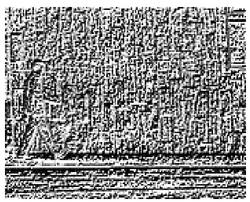
Spatio-temporal extension of textures

- Used three orthogonal plane (TOP) method [Zhao and Pietikainen, 2007]
- Encodes texture in XY, XT and YT planes (shape + space-time transition)
- Final feature vector is a concatenation of all planes:

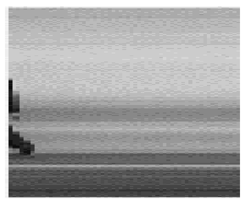
$$H = \{\tilde{h}^{XY}, \tilde{h}^{XT}, \tilde{h}^{YT}\}$$

Video in XY plane

LBP (XY)

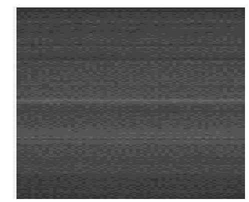


Video in XT plane



LBP (XT)

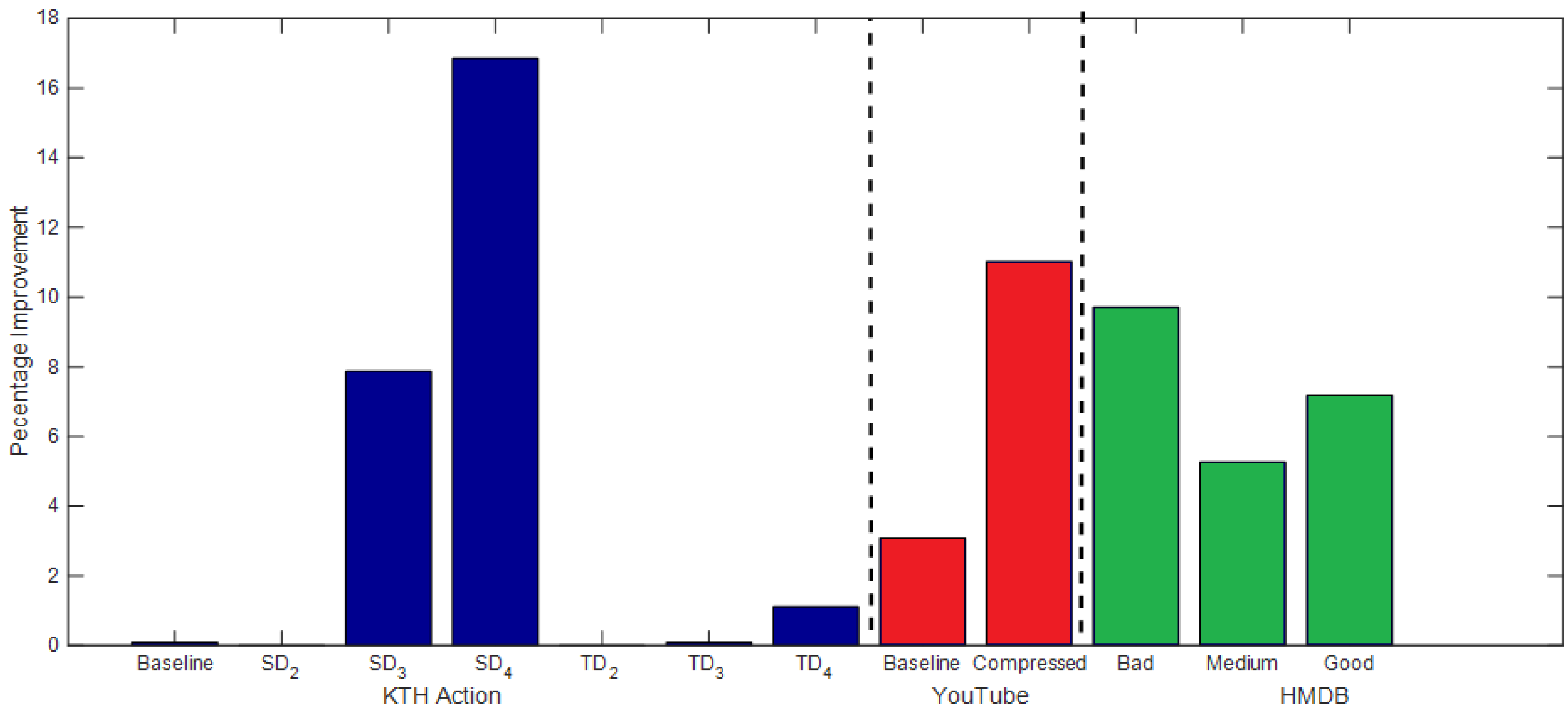
Video in YT plane



LBP (YT)

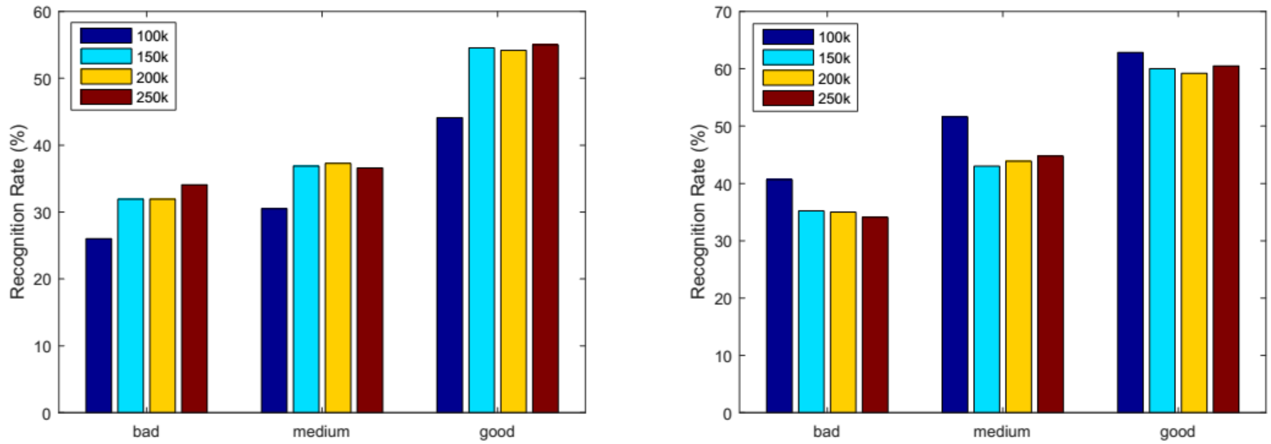
- Notation of textural features after TOP extension
  - $LBP - TOP\ P_{XY}P_{XT}P_{YT}R_{XY}R_{XT}R_{YT}$ (P is neighbourhood pixels and R is radius from centre in XY, XT and YT planes)
  - $LPQ - TOP\ W_x W_y W_t$ (W rectangular neighbourhood at each pixel position on XY, XT and YT planes)
  - $BSIF - TOP_{l,n}$ (rectangular filter l and representation bit size n at each pixel position on XY, XT and YT planes)
- Settings used for feature extraction
  - $LBP - TOP_{8,8,8,2,2,2}$ [Mattivi and Shao 09], $LPQ - TOP_{5,5,5}$, $BSIF - TOP_{9,12}$

Does textures also help good quality videos?



Performance improvement of iDT (FV) + BSIF-TOP over iDT (FV)

How feature sampling affects the performance?



(a) STIP+BSIF-TOP (b) iDT+BSIF-TOP

- More features more performance (only in case of STIP), not iDT!!
  - Feature ambiguity in codebook

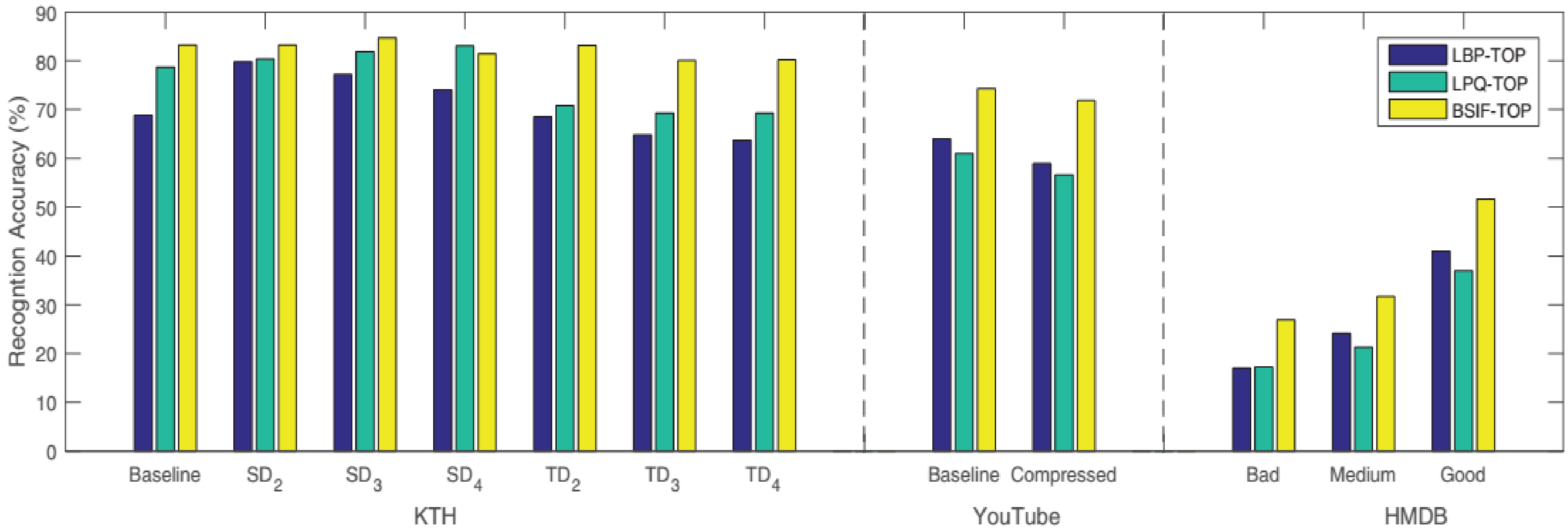
Computational Cost

- Comparison is performed on a sample video of 240x320 frame size and 246 frames (30 *fps*)
- Run-time was performed on Intel Core i7 3.60 GHz processor with 24GB RAM

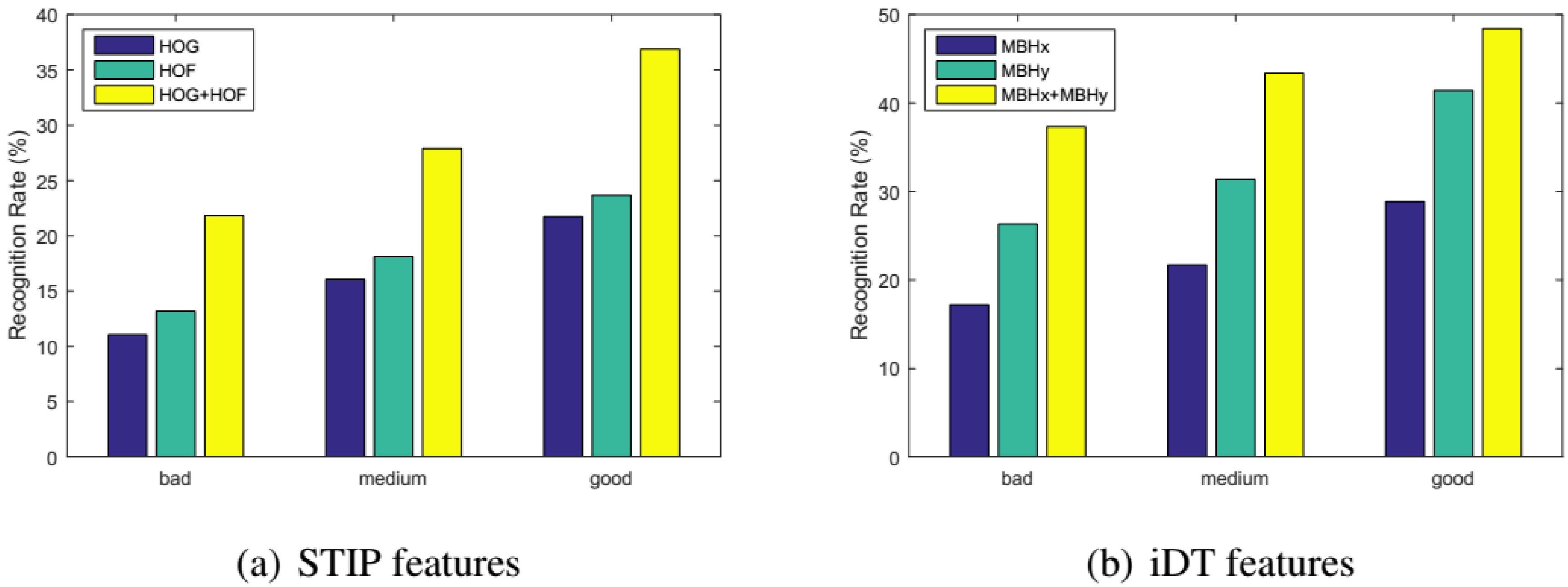
	STIP	iDT	LBP-TOP	LPQ-TOP	BSIF-TOP
Time per frame(in sec.)	0.156	0.203	1.230	0.041	0.051

Performance of various features (detection+description)

Performance of Textures



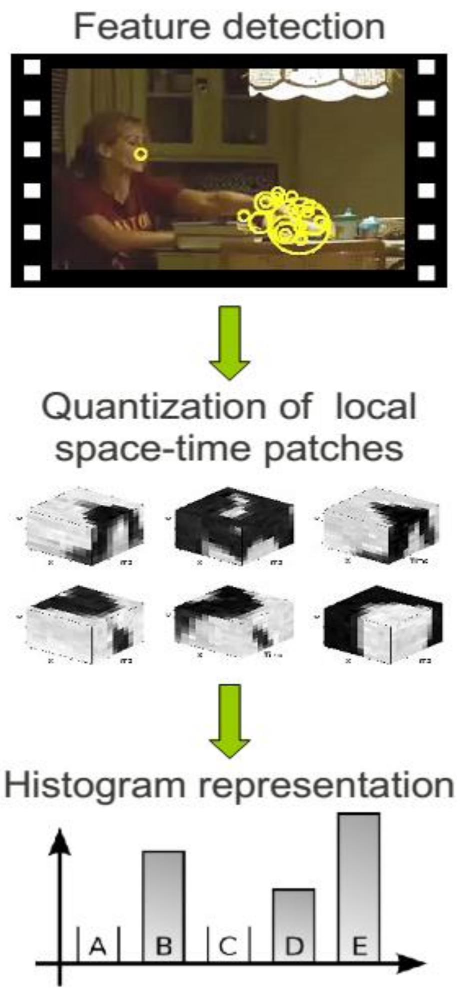
Performance of shape-motion features



Performance of various shape-motion features

Vector Quantization (VQ)

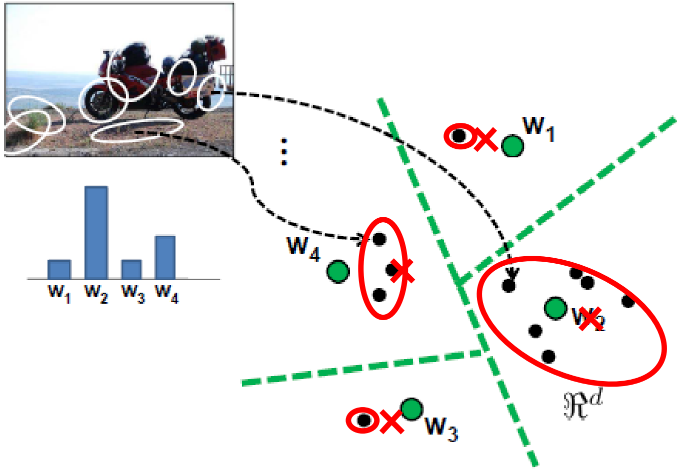
- Detection and description of local space-time features
- Codebook generation via clustering of training features (e.g., k-means, k=4000)
- Representation with occurrence histogram
  - Each feature is assigned to its closest cluster center (visual word)
- Classification of histograms (e.g., SVM with χ^2-kernel)



Fisher Vector (VQ)

•*Bag of Visual Words* is only about **counting** the number of local descriptors assigned to each Voronoi region

- Why not including **other statistics**? For instance:
  - mean of local descriptors
  - (co)variance of local descriptors

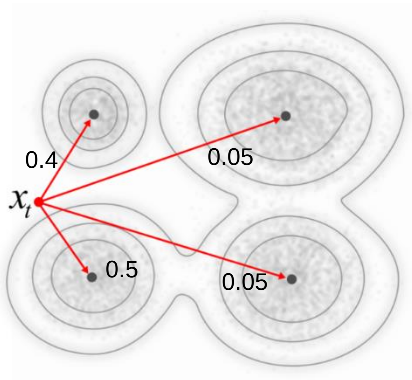


http://www.cs.utexas.edu/~grauman/courses/fall2009/papers/bag_of_visual_words.pdf

The Fisher vector

Relationship with the BOV

- FV formulas:
  - gradient wrt t



$$\boxed{\frac{1}{T}\sum_{t=1}^{T}\gamma_t(i)}$$

$\rightarrow$ **soft BOV**

- gradient wrt to μ and σ

$$\boxed{\begin{aligned}\mathcal{G}^X_{\mu,i} &= \frac{1}{T\sqrt{w_i}}\sum_{t=1}^{T}\gamma_t(i)\left(\frac{x_t-\mu_i}{\sigma_i}\right)\\ \mathcal{G}^X_{\sigma,i} &= \frac{1}{T\sqrt{2w_i}}\sum_{t=1}^{T}\gamma_t(i)\left[\frac{(x_t-\mu_i)^2}{\sigma_i^2}-1\right]\end{aligned}}$$

$\gamma_t(i)$ = soft-assignment of patch t to Gaussian i

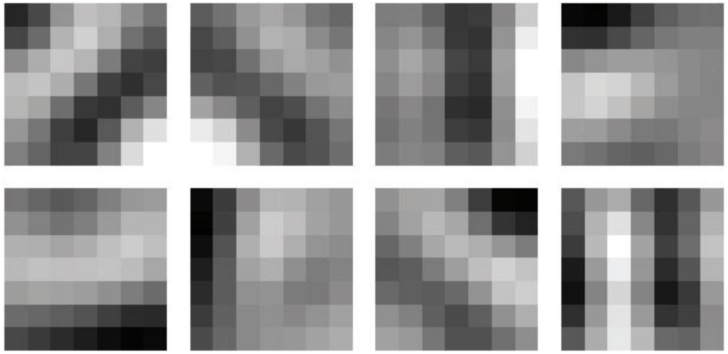
$\rightarrow$ compared to BOV, include **higher-order statistics**

- Let us denote: D = feature dim, N = # Gaussians
  - BOV = N-dim
  - FV = 2DN-dim

Perronnin and Dance, “Fisher kernels on visual categories for image categorization”, CVPR’07.

BSIF filter generation using ISA

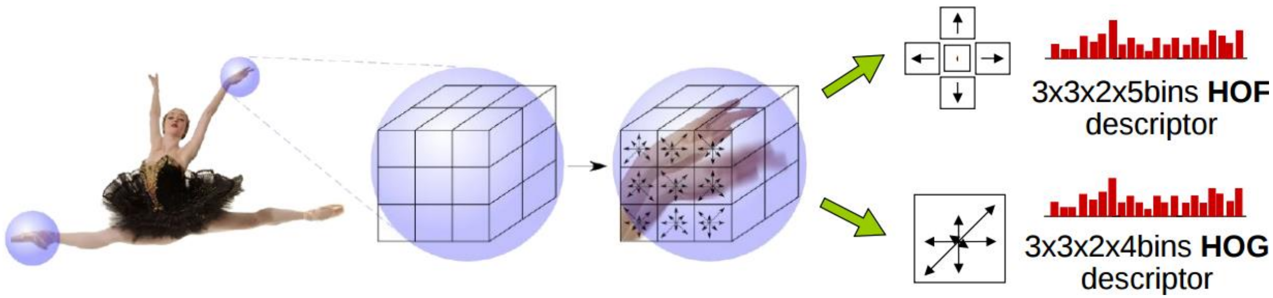
- A training set of image patches randomly sampled from natural images
- Patches are first made zero-mean and keep only n first PCs
- PCs are further divided by their standard deviation to get whitened data samples
- Use principal components algorithm [Hyvarinen & Oja, 2000] to estimate ICA filters



Learnt filters of size 9×9

A. Hyvarinen and E. Oja. Independent component analysis: algorithms and applications. Neural Networks, 2000

Spatio-temporal shape-motion feature encoding



= 90 bins

= 72 bins

$\sigma(x) = 3, \sigma(y) = 3$ and $\tau(x) = 2$

Constant Rate Factor (CRF) – x264

- Constant Rate Factor (CRF) is the default quality setting for x264 encoder
- CRF value distribution: 

- Keeps up a constant quality by compressing every frame of the same type the same amount.
  - maintaining a constant QP (quantization parameter) - how much information to “throw away” from a given block of pixels.
- X264 FFMEG does takes motion into account (compress different frames by different amounts)
- We used FFMPEG x264 video encoder

SVM Multi-Class Classification

- A SVM is a binary classifier, that is, the class labels can only take two values: ± 1.
- Many real-world problems, however, have more than two classes (e.g. optical character recognition).

One Versus the Rest: To get M-class classifiers, construct set of binary classifiers $f^1, f^2, \ldots, f^M$, each trained to separate one class from rest.

Combine them to get a multi-class classification according to the maximal output *before* applying the sgn function.

$$\underset{j=1\ldots M}{\operatorname{argmax}}\, g^j(x), \text{ where } g^j(x) = \sum_{i=1}^{m} y_i \alpha_i^j k(x, x_i) + b^j.$$

SVM Multi-Class Classification (cont.)

- Recall: $g^j(x)$ returns a signed real-valued value which can be interpreted as the distance from the separation (hyper)plane to the point x.
- Value can also be interpreted as a confidence value. The larger the value the more *confident* one is that the point x belong to the positive class.
- Hence, assign point x to the class whose confidence value is largest for this point.