Leveraging Textural Features for Recognizing Actions in Low Quality Videos

Saimunur Rahman, John See, Chiung Ching Ho

Centre of Visual Computing, Faculty of Computing and Informatics
Multimedia University, Cyberjaya 63100, Selangor, Malaysia

RoViSP 2016, Penang, Malaysia

Visual human actions

- Human actions: major visual events in movies, news, ...
- Low quality videos: low frame resolution, low frame rate, compression artifacts, motion blurring

- We recognize human actions from low quality videos
- Leverage textures with shape and motion features to improve action recognition form low quality videos.

Motivation

- Recognizing human actions from video is of central importance due to its large real-world application domain:
  - surveillance, human computer application, video indexing etc.
- Many methods have been proposed in recent years but majority are focused on high quality videos that offer fine details and strong signal fidelity.
  - not suitable for real-time and lightweight applications
- Current methods are not designed for processing low quality videos.

Summary of Approach

- Detect space-time patches by feature detector and describe using shape and motion descriptor.
- Calculate textural features from entire space-time volume.
- Combine shape, motion and textural features to improve performance.

Summary of Contribution

- Propose textural features to alleviate the limitation of shape and motion features.
- Use BSIF-TOP as a textural feature descriptor for action recognition in low quality videos.
- Evaluate various textural features on low quality videos.

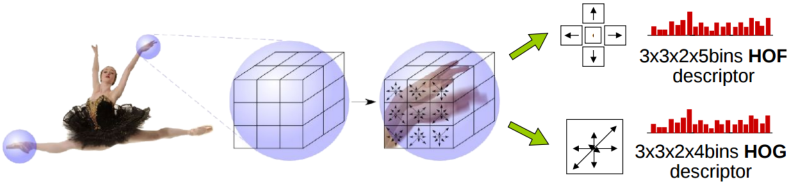
Related Work

- Shape and motion features
  - Space-Time Interest Points [Laptev et al'05]
  - Dense Trajectories [Wang et al.'11]
- Textural features
  - LBP-TOP [Kellokompu et al'09]
  - Extended LBP-TOP [Mattvi and Shao'09]
- Similar approaches
  - Joint Feature Utilization [Rahman et al'15, See and Rahman'15]

Outline

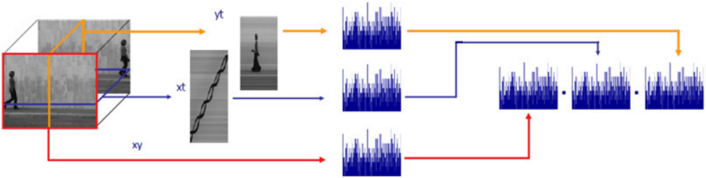
Shape and Motion Feature Representation

- Spatio-temporal interest points are detected by Harris3D detector [Laptev'05].
- Description of 3D patch around IPs using HOG and HOF [Laptev'08].
  - HOG - histogram of oriented gradients (encodes shape)
  - HOF - histogram of optical flow (encodes motion)



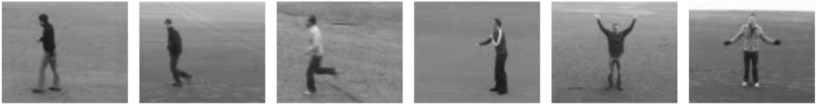
Textural Feature Representation

- Three types of textural features are calculated form entire space-time volume:
  - LBP - Local Binary Pattern [Zhao et al.'08].
  - LPQ - Local Phase Quantization [Zhao et al.'08].
  - BSIF - Binarized Statistical Image Features [Kannala and Rahtu'12].
- To obtain dynamic textures we apply three orthogonal plane (TOP) technique [Zhao et al.'08].
  - Features are calculated from XY, XT and YT plane of space-time volume (XYT).



Dataset : KTH Action [Schüldt et al'04]

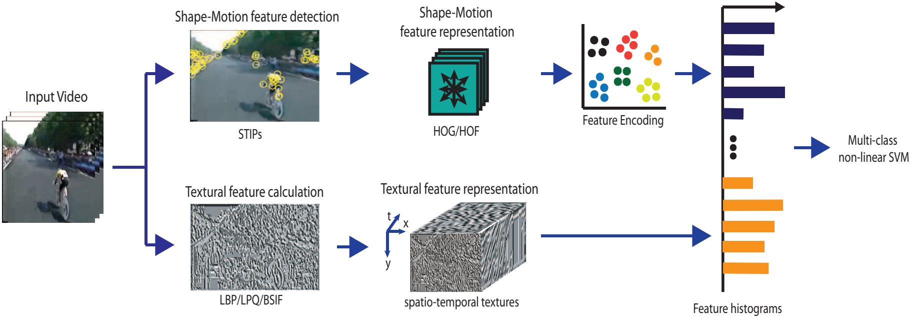
- Total 599 videos captured in a controlled environment.
- 6 action classes performed by 25 actors in 4 different scenarios.
- Sampling rate: 25 fps, Resolution: 160×120 pixels.
- Evaluation protocol: original experimental setup by authors.
- Six downsampled versions were cerated (3 spatial (SD_α) and 3 temporal (SD_β))
  - We limit $\alpha, \beta = \{2, 3, 4\}$, where α, β denotes spatial and temporal downsampling to half, one third and one fourth of the original resolution or frame rate respectively.



Dataset : HMDB51 [Oh et al'11]

- Total 6,766 videos of 51 action classes collected from movies or YouTube.
- Videos are annotated with a rich set of meta-labels including quality information
  - three quality labels were used, i.e. 'good', 'medium' and 'bad'.
- Evaluation protocol: three training-testing split by authors.
- We use the split specified for training, while testing is done using only videos with 'bad' and 'medium' labels; for clarity, we denote them as **HMDB-BQ** and **HMDB-MQ** respectively.

Evaluation Framework



Experimental Results: KTH dataset

- Performance (average accuracy over all class) comparison:

Method	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4
HOG/HOF [6]	83.33	76.39	65.74	86.11	81.94	76.85
HOG+HOF [9]	84.26	80.09	75.46	87.04	80.09	81.48
HOG+HOF + LBP-TOP [10]	87.41	80.74	77.69	87.87	82.50	80.37
HOG+HOF + LPQ-TOP	88.15	81.30	78.52	87.50	81.85	80.00
HOG+HOF + BSIF-TOP	**89.07**	**85.00**	**81.67**	**88.52**	**87.04**	**84.91**

- Best method: **HOG+HOF+BSIF-TOP**
- Spatially downsampled videos are highly benefited by textural features.
- BSIF-TOP outperform other textural features.

Experimental Results: HMDB51 dataset

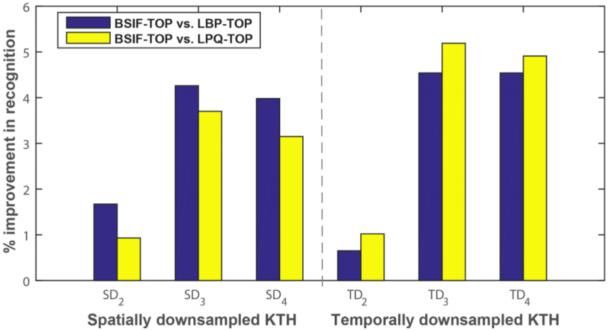
- Performance (average accuracy over all class) comparison:

Method	HMDB-BQ	HMDB-MQ
HOG/HOF [8]	17.18	18.68
C2 [8]	17.54	23.10
HOG+HOF [9]	21.71	23.68
HOG+HOF + LBP-TOP [10]	20.80	24.20
HOG+HOF + LPQ-TOP	23.89	28.36
HOG+HOF + BSIF-TOP	**32.46**	**37.14**

- Best method: **HOG+HOF+BSIF-TOP**
- Texture vastly improve the performance of both ‘Bad’ and ‘Medium’ quality videos.
- BSIF-TOP outperform other textural features.

Experimental Results: BSIF-TOP vs. other textures

- Performance improvement by BSIF-TOP over LBP-TOP and LPQ-TOP when aggregated with HOG+HOF:



- LPQ-TOP is better for spatially downsampled videos.
- LBP-TOP is better for temporally downsampled videos.
- Using BSIF-TOP, HMDB-LQ and HMDB-MQ results improves to almost double of baseline.

Experimental Results: Computational Complexities

- Computational cost (feature detection/calculation + quantization time) of various feature descriptors:

	HOG+HOF	LBP-TOP	LPQ-TOP	BSIF-TOP
Time (in sec.)	13.76	47.57	2.48	5.25

- Runtime reported using a Core i7 3.6 GHz 32GB RAM machine.
- All test run on a sampled video from KTH-SD_2 dataset consist of 656 frames.
- Ranking of descriptors in terms of speed:
  - LPQ-TOP > BSIF-TOP > HOG+HOF > LBP-TOP.

Conclusion

- We leveraged on textural features to improve the recognition of human actions in low quality video clips.
- Considering that most current approaches involved only shape and motion features, the use of textural features is a novel proposition that improves the recognition performance by a good margin.
- BSIF-TOP offers a significant leap of around **16%** and **18%** on the KTH-SD_4 and HMDB-MQ datasets respectively, over their original baselines.
- In future, we intend to extend this work towards a larger variety of human action datasets.
- It is also worth designing textural features that are more discriminative and robust towards complex backgrounds.

Acknowledgement

This work is supported, in part, by MOE Malaysia under Fundamental Research Grant Scheme (FRGS) project FRGS/2/2013/ICT07/MMU/03/4.

Thank You!

Q & A