On the Effects of Low Quality Video in Human Action Recognition

John See, Saimunur Rahman

Centre of Visual Computing, Faculty of Computing and Informatics, Multimedia University, Malaysia

Motivations

- Lack of action recognition works that deal with the problem of low quality videos
- Popular space-time feature descriptors do not generalise well when details are less accurate

Scope

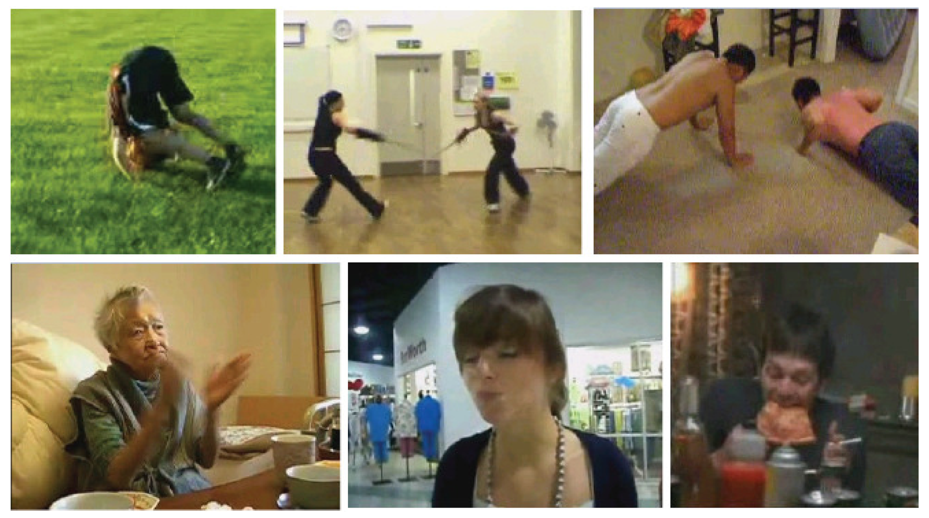
Low Quality: Focus is on videos that are poor in the aspect of resolution (spatial sampling), frame rates (temporal sampling), and compressed videos affected by motion blurring and compression artifacts.

Contributions

- Investigate the performance of popular representations for action recognition when video quality is poor
- Propose the use of spatio-temporal texture features to complement shape and motion
- Report extensive evaluation on two benchmark action datasets – KTH and HMDB51

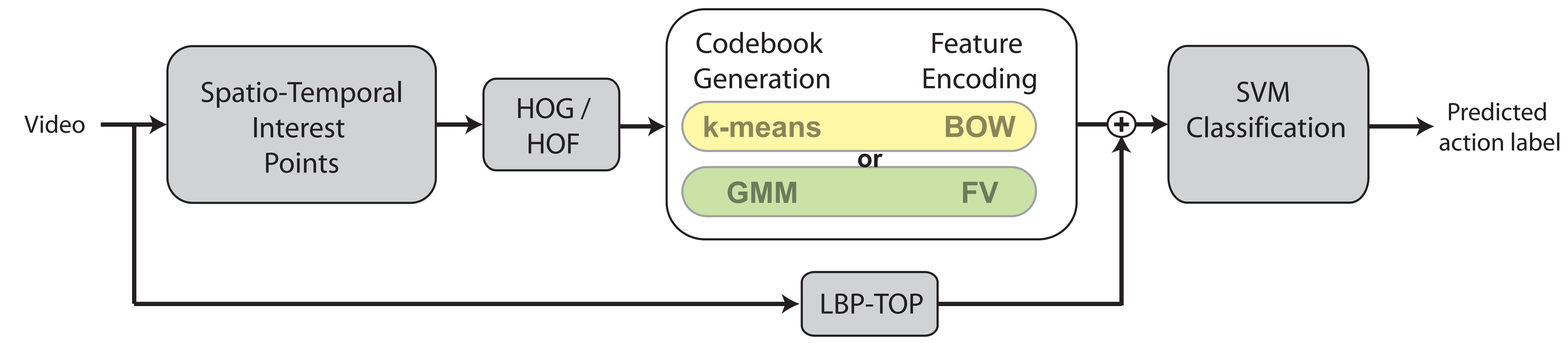
Datasets

KTH *(small-scale, simple backgrounds, downsampled)*



HMDB51 *(large-scale, complex backgrounds, motion blur, compression artifacts)*

Proposed Framework



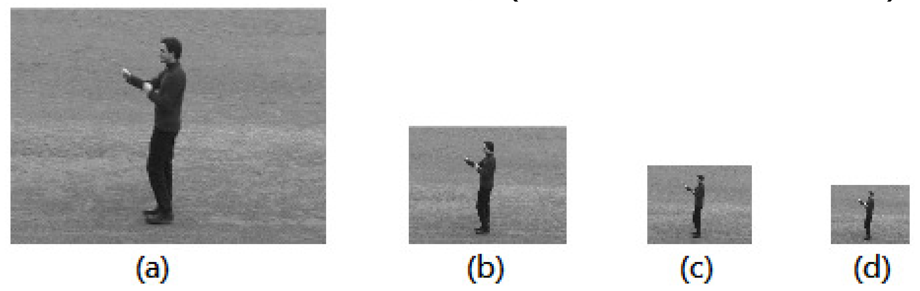
Downsampled KTH Results

Recognition accuracy (%)

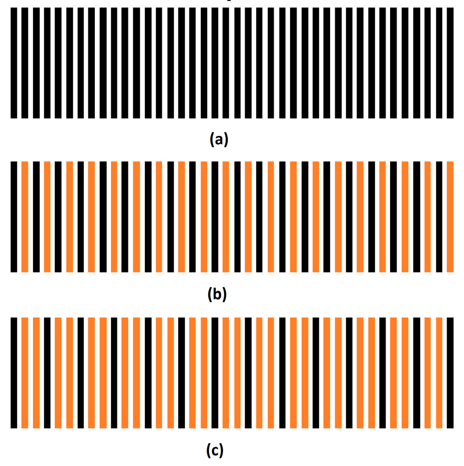
Method	BOW (V=4000)						FV (K=256)					
	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4
HOG	76.85	66.20	55.56	80.09	76.85	75.46	75.00	69.44	55.09	86.57	81.94	84.26
HOG+LBP-TOP	80.56	73.61	76.39	80.56	75.46	74.54	79.63	76.85	75.93	85.19	83.80	79.17
HOF	88.89	82.41	76.39	83.80	75.46	72.22	87.50	82.87	76.38	85.19	81.94	76.85
HOF+LBP-TOP	**89.35**	**85.65**	**84.26**	83.80	80.56	78.70	**88.43**	**82.87**	**81.94**	86.11	83.80	78.70
HOGHOF	83.33	76.39	65.74	86.11	81.94	76.85	86.11	80.09	64.35	88.43	84.26	82.87
HOGHOF+LBP-TOP	86.11	77.31	77.31	**89.35**	**85.65**	**81.94**	87.04	82.41	78.70	**90.28**	**85.19**	**84.72**

Video Downsampling

Spatial Downsampling (SD_2, SD_3, SD_4)



Temporal Downsampling (TD_2, TD_3, TD_4)



Methods

- Spatio-temporal Interest Points: **Harris 3D**
- Local Shape & Motion Descriptors: **HOG, HOF**
- Local Textural Descriptor: **LBP-TOP**
- Codebook Generation: 1) **Bag-of-Words (BoW)**, 2) **Fisher Vector (FV)**
- Classification: Multi-class **SVM** with χ^2-kernel

Analysis & Discussions

KTH

- Spatial resolution $\downarrow$: HOF+LBP-TOP limits $SD_2 \rightarrow SD_4$ to only $\sim 5\%$ drop
- Temporal frame rate $\downarrow$: HOGHOF+LBP-TOP limits $TD_2 \rightarrow TD_4$ to only $\sim 6\%$ drop

HMDB51

- HMDB Overall: Out of 51 classes, 20 improved, 9 drop, rest unchanged.
- HMDB-MQ: $> 60\%$ improvement over baseline
- HMDB-BQ: $> 70\%$ improvement over baseline

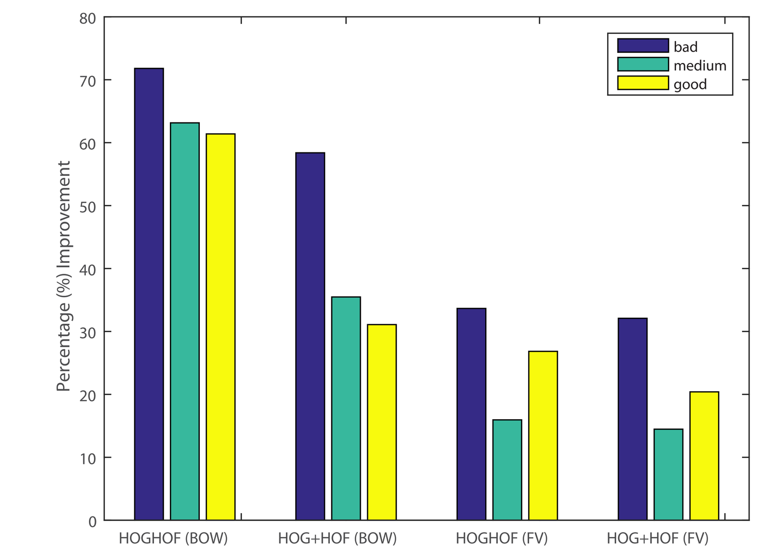
Codebook generation

- Random Sampling Size for training codebook: 200k descriptors (best empirically)
- Encoding methods: FV has no advantage over BoW when spatial resolution $\downarrow$, FV $>$ BoW for complex scenes (HMDB51)
- LBP-TOP has negligible effect on the complexity of codebook, i.e. $\ell_{LBPTOP} \ll \ell_{STIP}$ which is V for BoW, or $2DK$ for FV

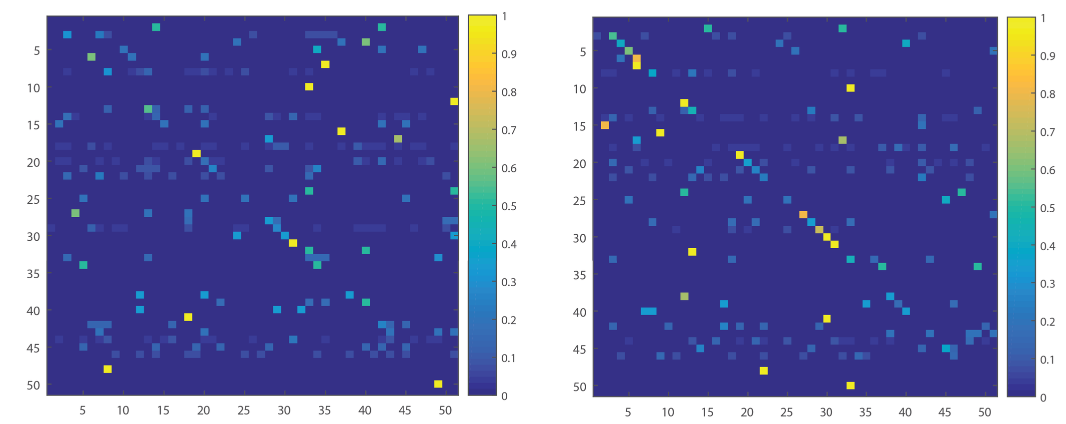
HMDB Low Quality Subset Results

Recognition accuracy (%)

Method	HMDB-BQ		HMDB-MQ	
	BoW	FV	BoW	FV
HOG+HOF	16.44	21.57	22.87	30.79
HOGHOF+LBP-TOP	23.48	**28.66**	28.32	33.94
HOG+HOF+LBP-TOP	**26.04**	28.49	**30.99**	**35.24**
HOGHOF (Baseline) [1]	17.18	-	18.68	-
C2 (Baseline) [1]	17.54	-	23.10	-
LBP-TOP [2]	17.00		24.11	



Percentage (%) of increment after textural features are considered



Confusion matrices for HOG+HOF (left) & HOG+HOF+LBP-TOP (right)

References

[1] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). HMDB: A large video database for human motion recognition. In *ICCV*, pages 2556–2563.

[2] Zhao, G. and Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. PAMI*, 29(6):915–928.

Acknowledgements

This work is supported, in part, by the Ministry of Education, Malaysia under FRGS project FRGS/2/2013/ICT07/MMU/03/4

Contact Information

URL: http://pesona.mmu.edu.my/~johnsee

Email: johnsee@mmu.edu.my