ICASSP 2016 Shanghai, China

Spatio-Temporal Mid-Level Feature Bank for Action Recognition in Low Quality Video

Saimunur Rahman, John See

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

MMU MULTIMEDIA UNIVERSITY

Motivations

- Very few works deal with the problem of action recognition in low quality videos
- Popular local space-time features (shape, motion) are ineffective when video quality deteriorates

- Textural features can complement well but produce indiscriminate features due to unrelated background motion and pixel-based artifacts

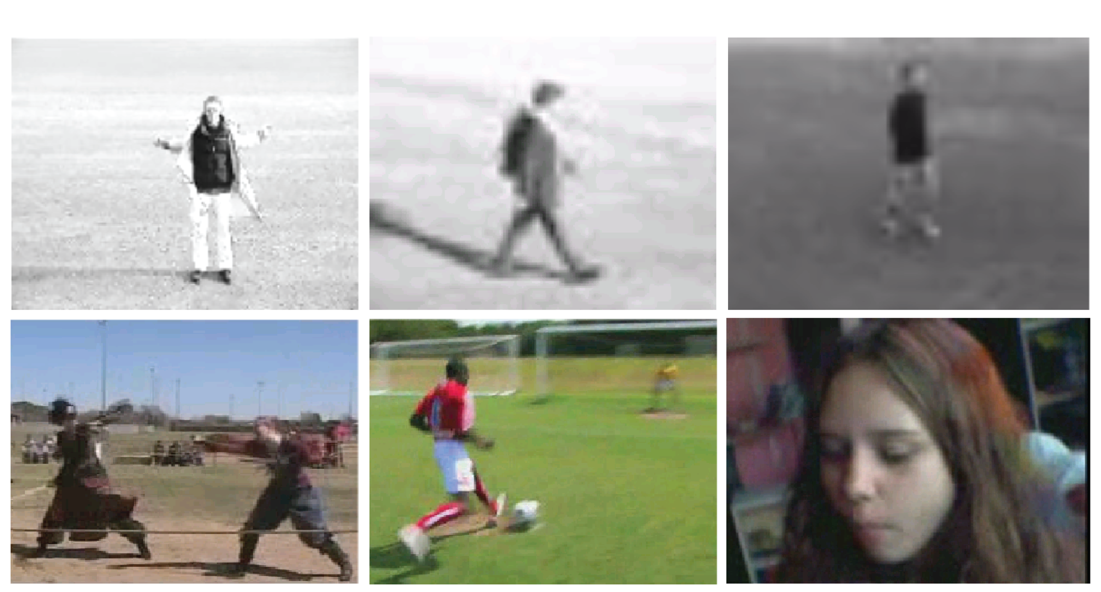
Scope

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

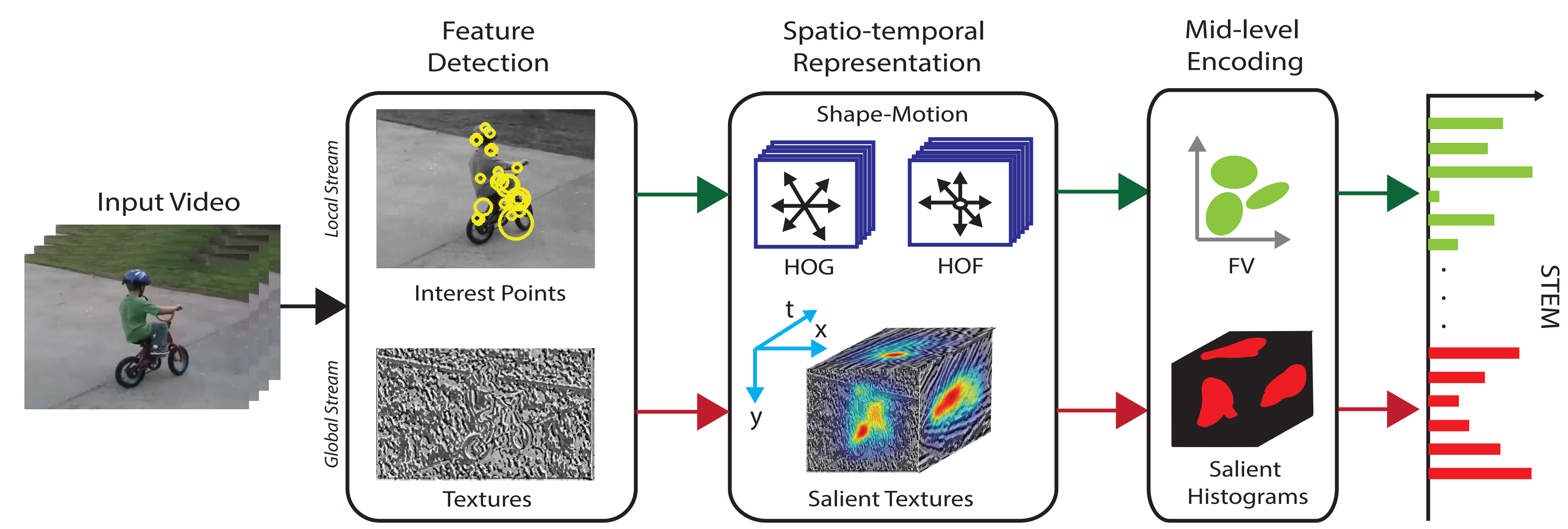
- A new spatio-temporal mid-level (STEM) feature bank for recognizing actions in low quality videos is introduced
- Features are detected at local and global *streams* to exploit the benefits of local shape-motion and global statistical patterns
- Salient textural histograms are extracted discriminately based on 3D salient patches

Datasets



Low quality samples from **KTH** and **HMDB51**

Proposed STEM Encoding Framework

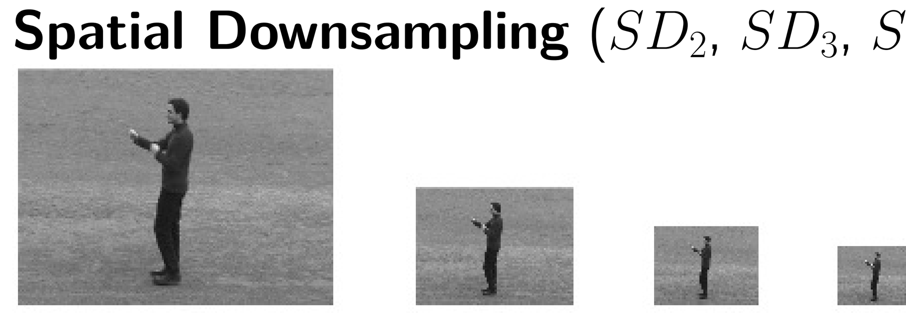


Results on Low Quality Versions/Subsets

Method	KTH (downsampled)						HMDB	
	SD_2	SD_3	SD_4	TD_2	TD_3	TD_4	BQ	MQ
HOG+HOF (BoW encoding) [3]	88.24	81.11	73.89	87.04	82.87	82.41	17.40	22.77
HOG+HOF	89.63	82.31	78.98	89.35	86.11	83.89	26.02	30.53
HOG+HOF+LBP-TOP	89.81	81.48	78.70	89.35	86.11	84.72	28.49	35.24
STEM (w/o salient textures)	89.35	82.87	79.72	89.63	87.41	84.63	33.78	38.76
STEM	88.52	**83.98**	**83.15**	**90.00**	**88.06**	**85.09**	**34.08**	**38.94**

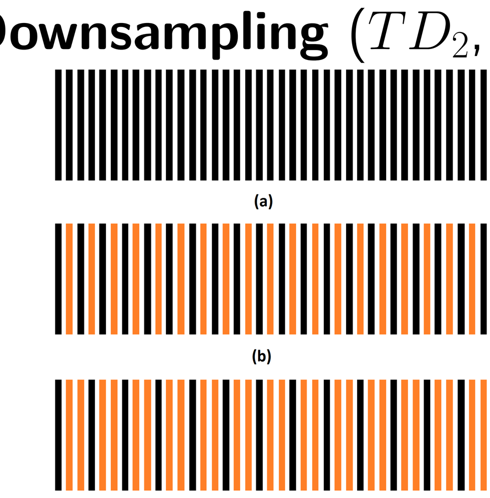
Video Downsampling

Spatial Downsampling (SD_2, SD_3, SD_4)



(a) (b) (c) (d)

Temporal Downsampling (TD_2, TD_3, TD_4)



(a) (b) (c)

Local Stream

- Spatio-temporal Interest Points: **Harris 3D**
- Local Shape & Motion Descriptors: **HOG, HOF**
- Encoded by **Fisher Vector (FV)**

Global Stream

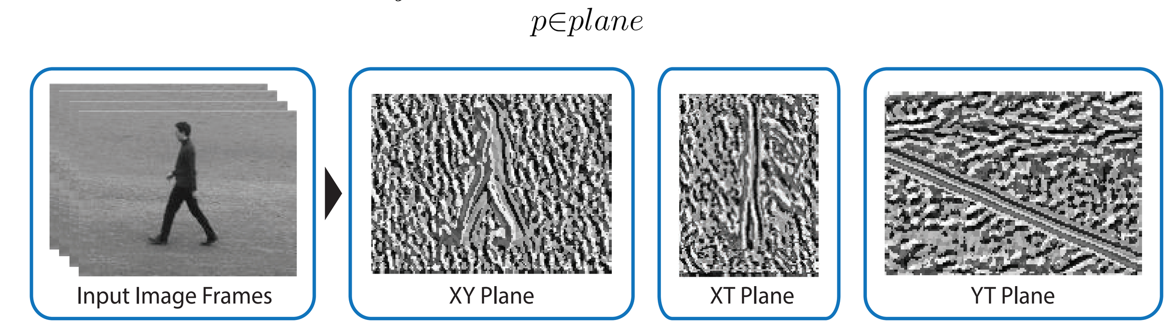
ST Textural Features

- Binarized statistical image features (BSIF) [2]
  - Filter response

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

  is thresholded at level zero to obtain binarized feature b_i
  - n number of filters produce n-bit binary code
  - TOP extended (XY, XT, YT planes)

$$h_j^{plane} = \sum_{p \in plane} \mathcal{I}\{b_i(p) = j\}$$



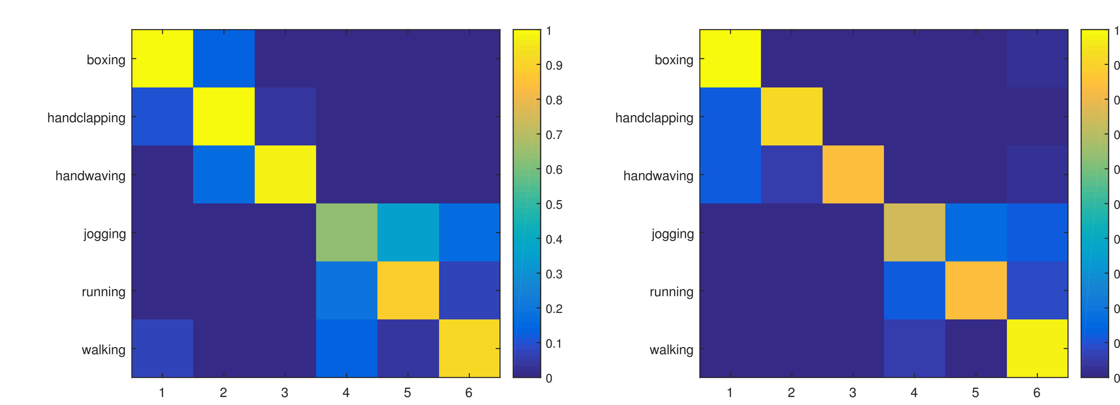
Salient Histograms

- Graph-based visual saliency (GBVS) [1]
  - 3 features maps: Contrast, orientation, flicker.
  - Saliency map $S_{i,j}$ is converted to binary saliency mask $Z_{i,j}$ by Otsu's method.
  - Applying saliency to the j-th bin of BSIF histogram yields

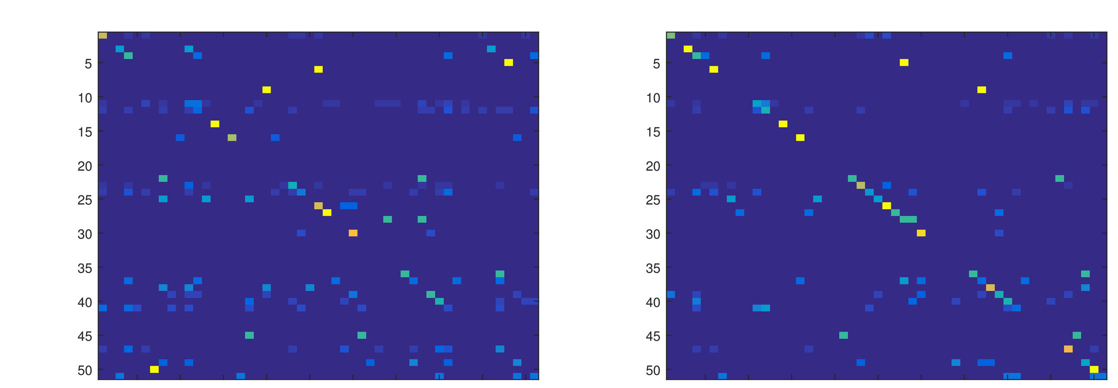
$$\bar{h}_j^{plane} = \sum_{p \in plane} \mathcal{I}\{\{b_i(p) = j\} \cap \{Z(p) = 1\}\}$$

Analysis & Discussion

- **KTH**: STEM outperforms other methods in all versions (except SD_2)
  - STEM is most robust under low spatial resolutions
  - STEM is increasingly stronger when quality drops – observe SD_4 and TD_4
- **HMDB51**: Accuracy ↑ by $\sim 8\%$ for BQ and MQ subsets
- **Multi-scale salient features**: Filter sizes dictate "scale" of information – using 3 scales {3, 9, 15} can increase performance by another 1–2%



Confusion matrices of KTH-SD_3 for STIP (left) & STEM (right)



Confusion matrices of HMDB for STIP (left) & STEM (right)

References

[1] Harel, J., Koch, C., and Perona, P. (2006). Graph-based visual saliency. In *NIPS*, pages 545–552.

[2] Kannala, J. and Rahtu, E. (2012). Bsif: Binarized statistical image features. In *ICPR*, pages 1363–1366.

[3] 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.

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