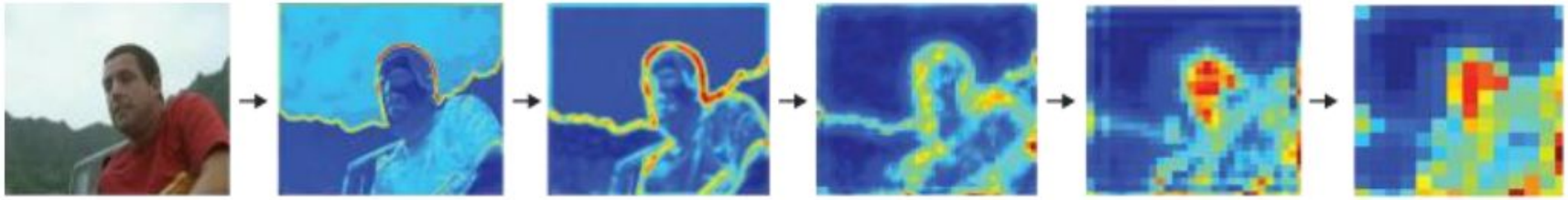


Deep CNN Object Features for Improved Action Recognition in Low Quality Videos

Saimunur Rahman, John See and Chiung Ching Ho

Visual Processing Laboratory
Multimedia University, Cyberjaya

At first, the overview of this talk

1. Introduction
2. Problem statement
3. Related Works
4. Proposed Method
5. Experimental Results
6. Conclusion

Introduction

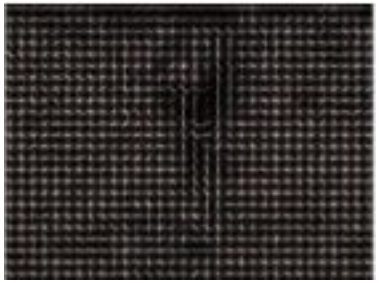
- Proposed a hybrid solution for activity recognition in low quality videos
  - Leverage both handcrafted and deep-learned features
- Achieved competitive results for low quality subsets of two publicly available datasets
  - Low quality version of UCF-11 [Liu et al. 2009]
  - Low quality subsets from HMDB51 [Kuehne et al. 2011]

Problem Statements

- Handcrafted features estimation is …
  - Lack robust image structure encoding
    - Highly dependent on image resolution
  - Mostly rely on local features
    - May miss important image region
- Leverage scene and objects
  - Use context of the action-of-interest

Original Frame

HOG Orgi. Res.



CRF 50

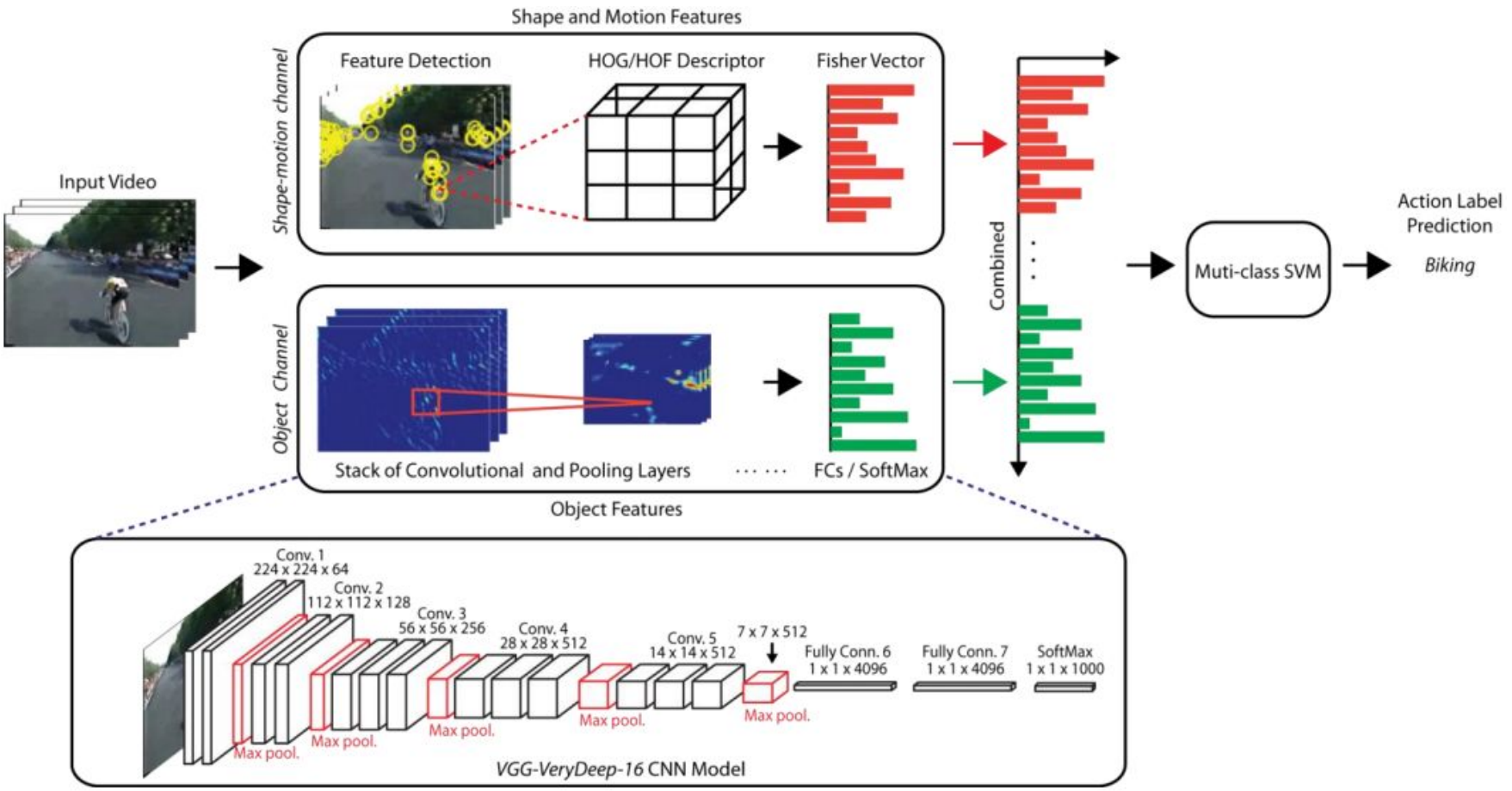
CRF 40

Related Works

- Handcrafted Features
  - **Detectors:** STIP [Laptev et al. 2003], Cuboid [Dollar et al. 2009], iDT [Wang et al. 2015] etc.
  - **Descriptors:** HOG/HOF [Laptev et al. 2003], MBH [Wang et al. 2011] etc.
- Deeply-learned features
  - **CNN based:** 3D-CNN [Karpathy et al. 2014],

    Two-stream CNN [Simonyan and Zisserman. 2014] etc.

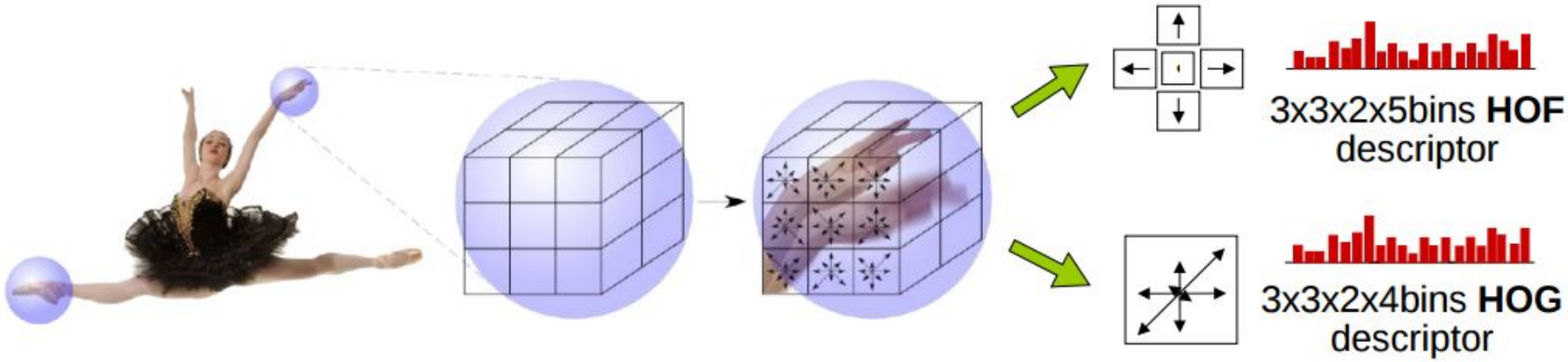
Proposed Framework



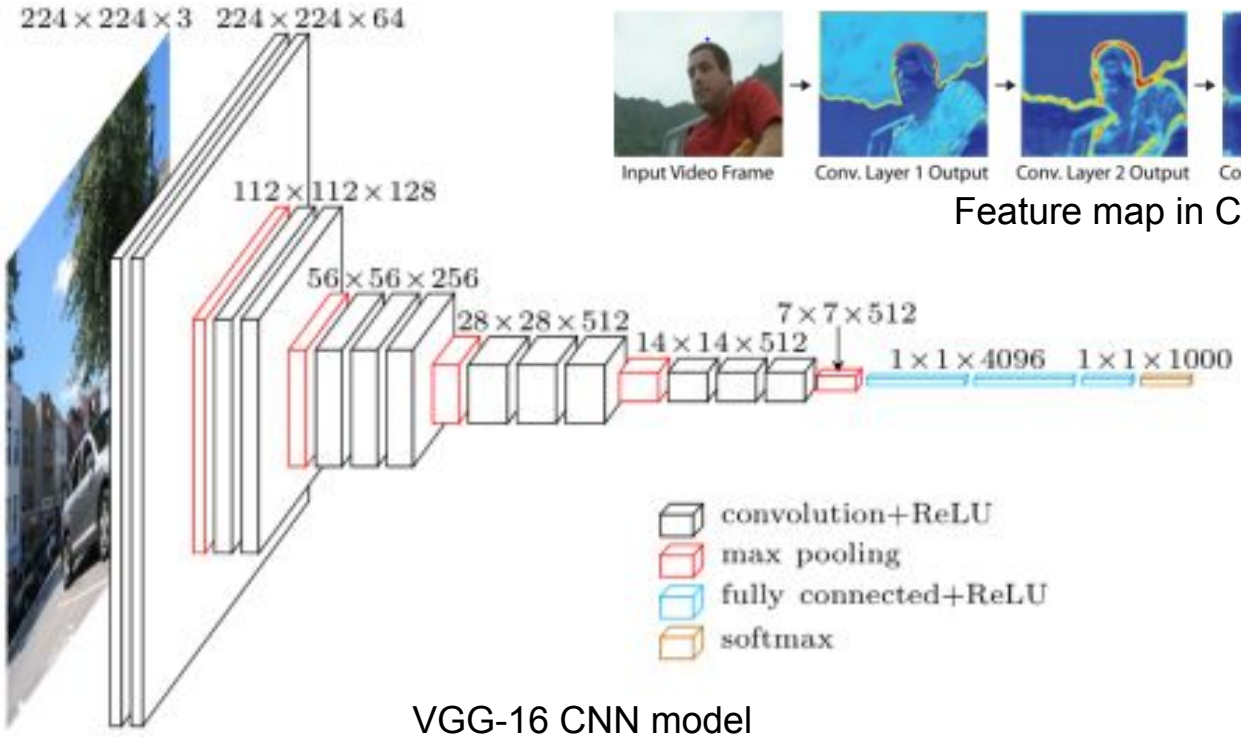
- **Shape-motion Channel:** Harris3D + HOG/HOF
- **Object Channel:** VGG-16 trained on ImageNet + FCs/SoftMax
- **Classification:** multi-class SVM + chi^2 homogeneous kernel

Shape-motion features

- STIP driven shape + motion features
  - **STIP detection:** Harris3D [Laptev and Linderberg. 2003]
  - **Shape feature:** Histogram of Oriented Gradients (HOG) [Laptev et al. 2008]
  - **Motion feature:** Histogram of Optical Flow (HOF) [Laptev et al. 2008]



Deep Object Features



VGG-16 CNN model

Feature map in Conv. Layers

- VGG16 very deep CNN model [Simonyan and Zisserman. 2014] trained on 1000 categories of ImageNet
- Not sufficient to describe frame-object level features with higher degree of discriminativeness
- Last Conv. layers offers more rich features (comparable with mid-level like features)
- **Deep Object Features:** FC6, FC7 and SoftMax

Datasets

- Two publicly available datasets
  - UCF-11 dataset
    - 11 action classes, 1600 videos, Video resolution: 320x240
    - Compressed with uniform CRF distribution: CRF 23-50
  - HMDB51 dataset
    - 51 action classes, 6766 videos
    - Quality-based test-train split: Good, Medium and Bad, Use Bad and Medium for test

Sample low quality videos

Class-specific CRF values for UCF-11: http://saimunur.github.io/YouTube-LQ-CRFs.txt

Experimental Result (Individual channel)

Table 2. Experimental Results of Shape and Motion Features on Various Low Quality Datasets

METHOD	Dim.	UCF-LQ	HMDB BQ	HMDB MQ
HOG	36864	63.57	8.15	10.40
HOF	46080	59.10	11.41	10.65
HOG+HOF	82944	**70.27**	**26.02**	**30.53**

Table 1. Experimental Results of Various Object Features on the Low Quality Datasets

METHOD	Dim.	UCF-LQ	HMDB BQ	HMDB MQ
Softmax	1000	77.42	23.31	30.46
FC6	4096	**83.54**	23.31	30.50
FC7	4096	81.33	28.41	38.02
FC6+FC7	8192	83.13	**31.99**	**39.63**
FC6+FC7+softmax	9192	83.08	31.98	**39.70**

Experimental Result (channel combined)

Table 3. Experimental Results of Combination of Shape, Motion and Object Features on Low Quality Datasets

METHOD	Dim.	UCF-LQ	HMDB BQ	HMDB MQ
HOG+FC6+FC7	45056	84.03	33.02	40.05
HOF+FC6+FC7	54272	85.16	32.80	40.41
HOG+HOF+FC6+FC7	91136	**86.34**	33.74	**40.55**
HOG+HOF+LBP-TOP[17]	85248	70.99	23.88	30.71
HOG+HOF+LPQ-TOP[23]	86016	71.65	25.02	30.75
STEM (w/o saliency)[24]	87040	75.04	33.78	38.76
STEM[24]	87040	77.50	**34.08**	38.94

Computational Complexity

Table 4. Computational Cost of Feature Extraction by shape-motion descriptors (feature detection+description) and *'VGG-VeryDeep-16'* object model

METHOD	Harris3D+ HOG+HOF (shape-motion)	VGG model (object)
Time per frame (sec.)	0.156	0.303

- Test Scenario
  - A video from bike_riding class of HMDB51
    - 240x320 pixels and 246 video image frames at 30 *fps*
  - Intel Core *i7* PC with 24GB memory

Conclusion and future work

- Proposed to use image-trained deep CNN model to obtain object features for video based activity recognition.
- Deep CNN features are proven to complement traditional shape-motion features, also HAR in LQ videos.
- Can be further improved by fine-tuning CNN model by action images.

Acknowledgements

- **FRGS** grant FRGS/2/2013/ICT07/MMU/03/4
- MMU Internal Conference Travel Grant

Thank You

Any Questions?