Semantic Video Classification by Fusing Multimodal High-Level Features

Olivier Nguyen, Yongqing Sun, Kyodo Sudo, Akira Kojima
NTT Media Intelligence Lab. NTT Corporation,

MIRU 2015, July 2015

1. Objective of Our Work

Semantic Video Classification

Automatic Video Classification by Combining High-level Features

2. Related Work

① Bank Representation

➢ Object Bank

[L. Li, 10], Object Bank models an imaged based on the objects that appear in it

➢ Action Bank

[S. Sadanand & J. Corso, 12], Action Bank uses action detectors to form the video representation

② Improved Densed Trajectories

[H. Wang, 13], Dense Trajectories samples dense points & tracks them from optical flow

③ Two-Stream Convolutional Networks for Action Recognition

[K. Simonyan, 14], Deep learning method that combines still-frames and motion

3. Overview

Main Steps of Our Method:

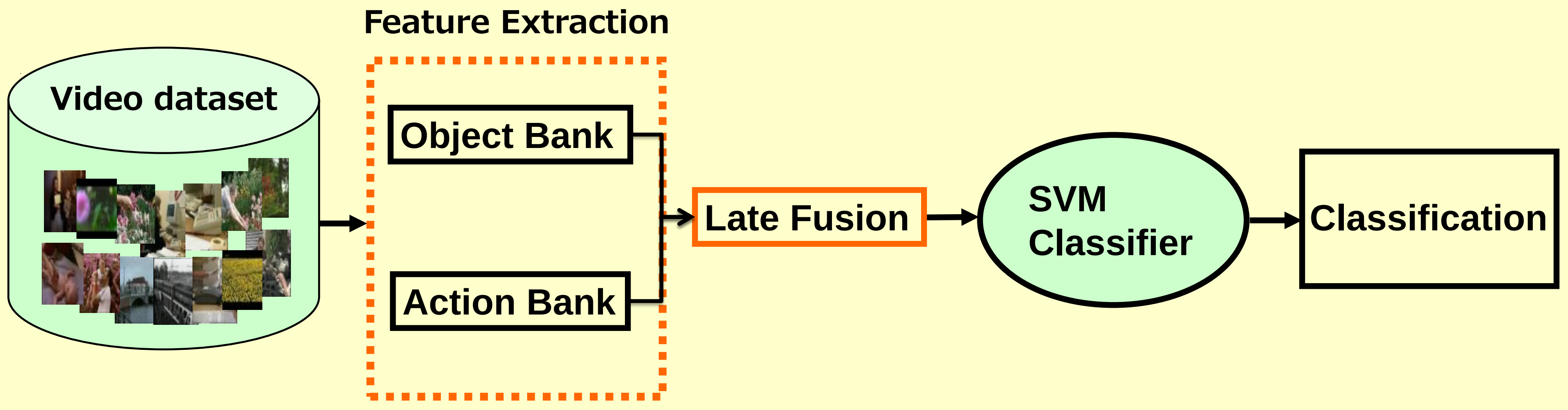
① Object Bank & Action Bank feature extraction

Feature vectors are mean-pooled across each detector to represent presence or absence of objects/actions

② Fusion of features

Using late fusion method of weighted averaging

③ Train an SVM classifier



Feature Extraction

Video dataset

Object Bank

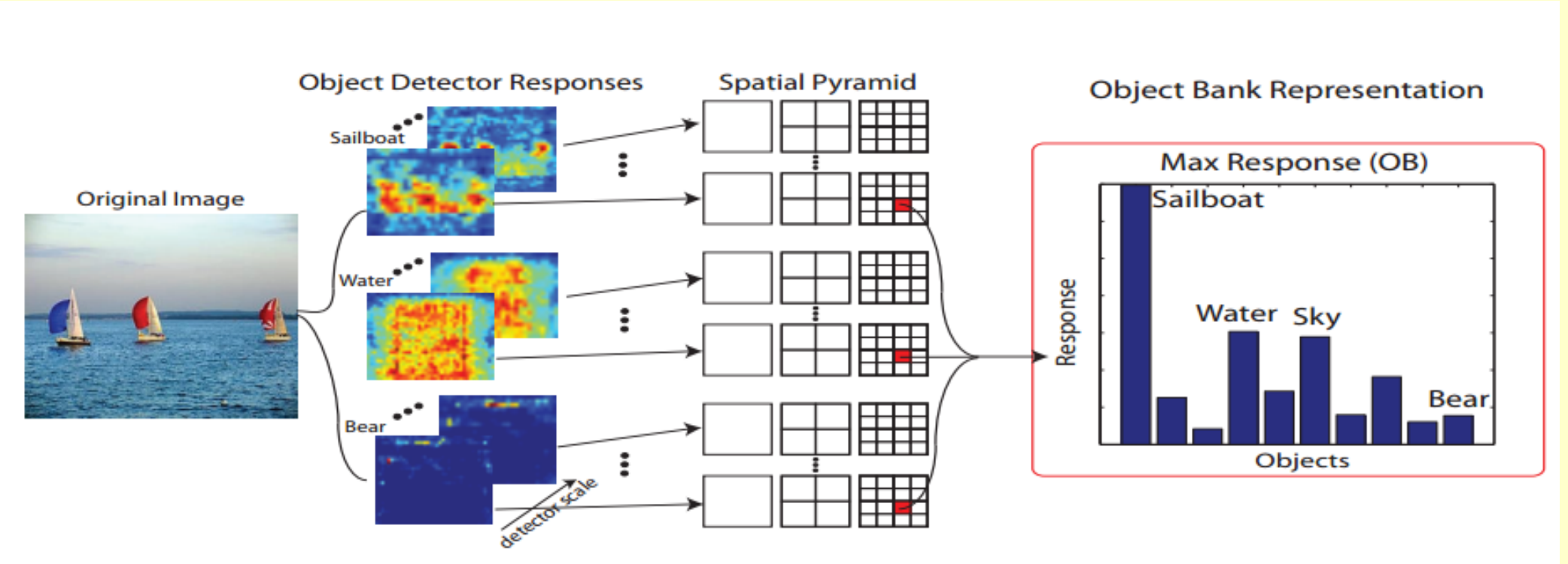
Action Bank

Late Fusion

SVM Classifier

Classification

Example: Images with similar low-level image information

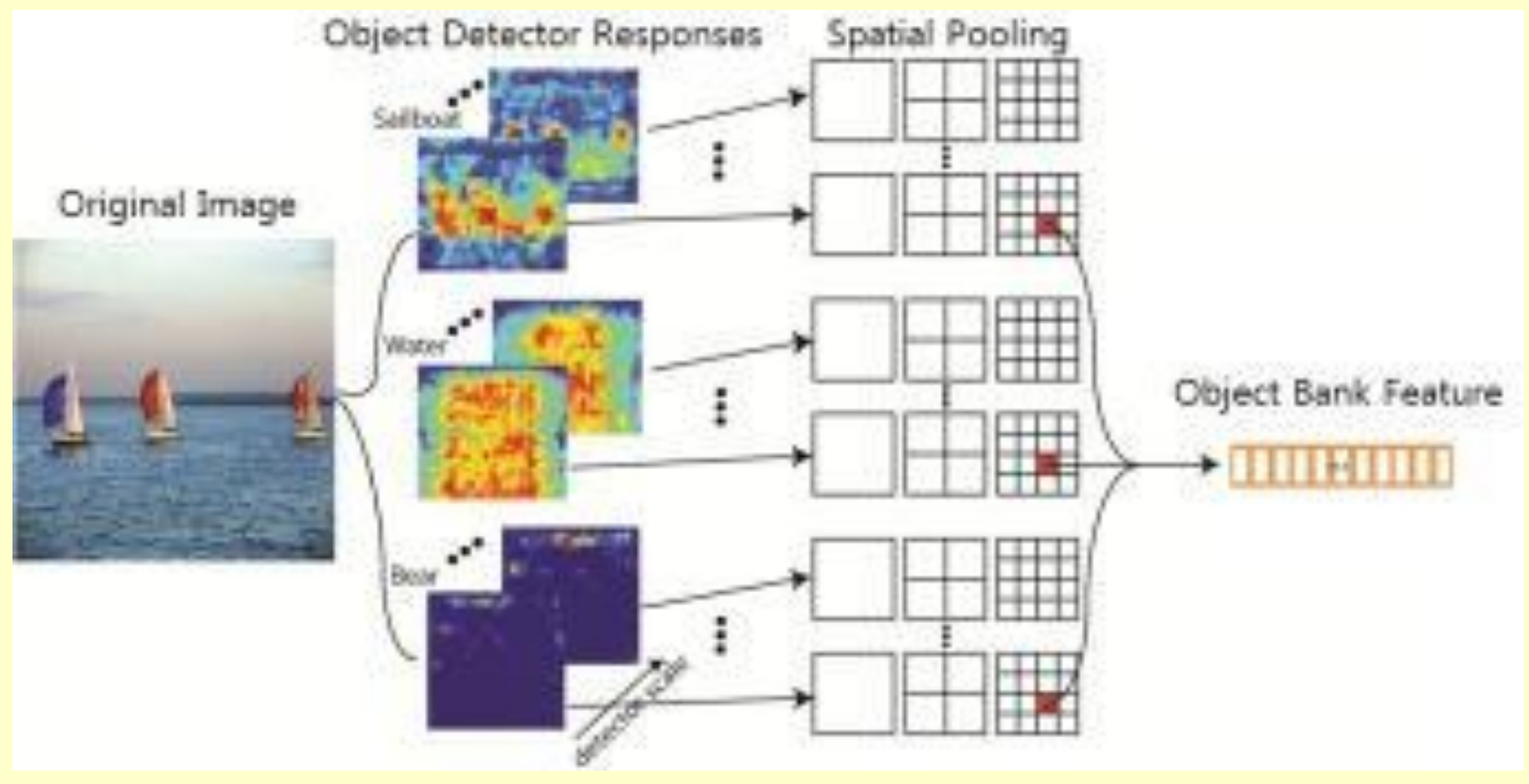


4. Details of Our Method

① Feature Extraction & Preprocessing

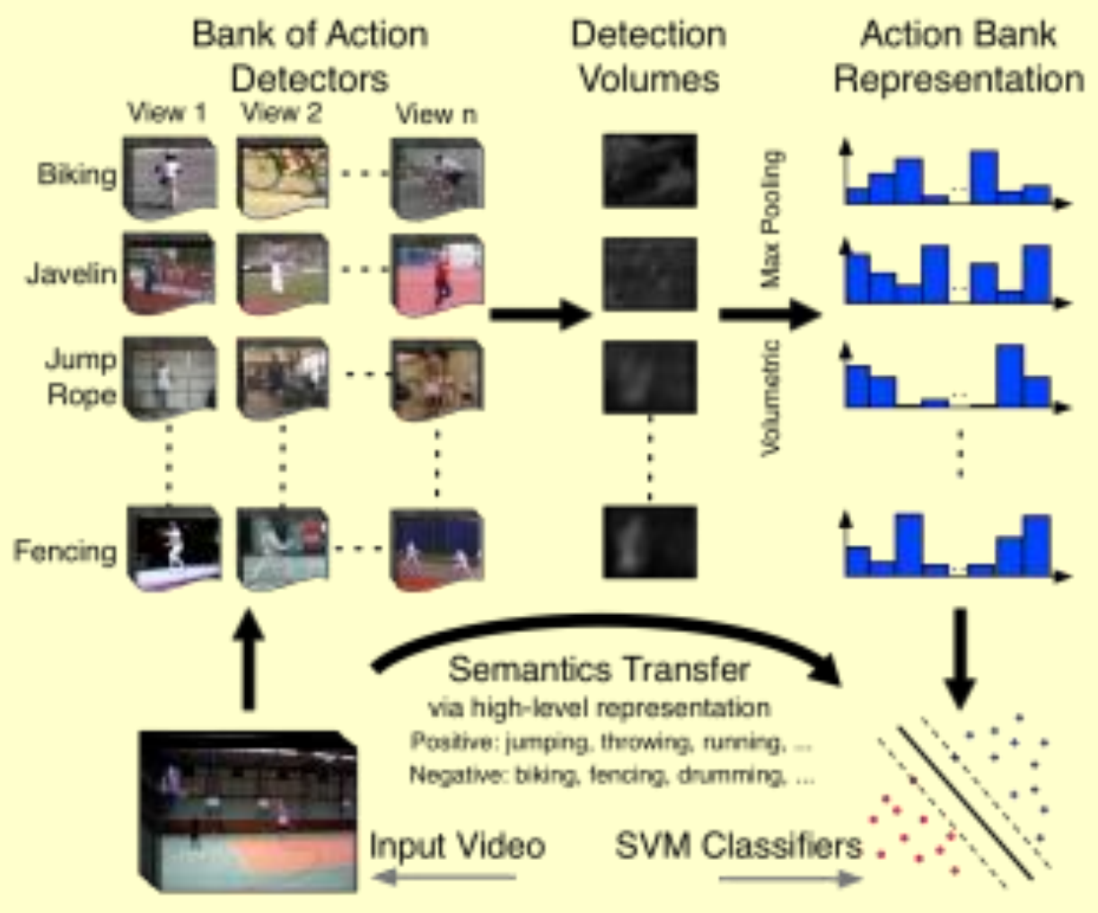
①.1: Object Bank

- ◆ **Key-frame are extracted from the raw input videos**
- ◆ **Images constituting the video are max-pooled on all dimensions**



①.2 : Action Bank

- ◆ **Features are extracted directly from the input videos**
- ◆ **Each detector is mean-pooled**



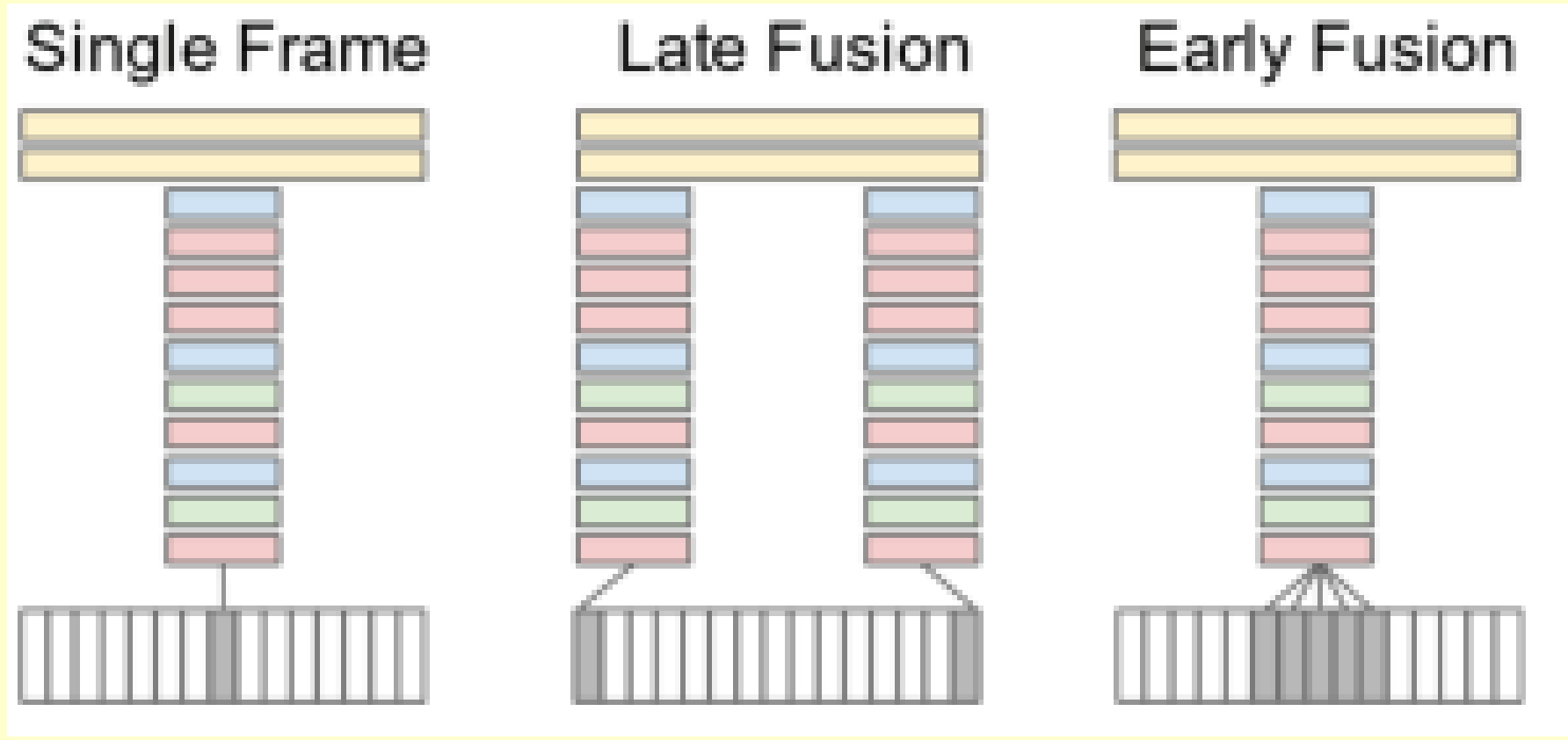
② Feature Fusion Methods

◆ **Weight Averaging (WA) as late fusion method to combine both Features**

$$p(c \mid x_1, \ldots, x_M) = \sum_{i=1}^{M} p(c \mid x_i)\alpha_i$$

c: video class, xi: individual feature, M: number of features, α: weight value

◆ **Weights α selected by exhaustive grid search**



③ Training of Classifiers

◆**SVM was used for classification with kernels and hyper-parameters selected through grid-search and cross-validation**

5. Experiment

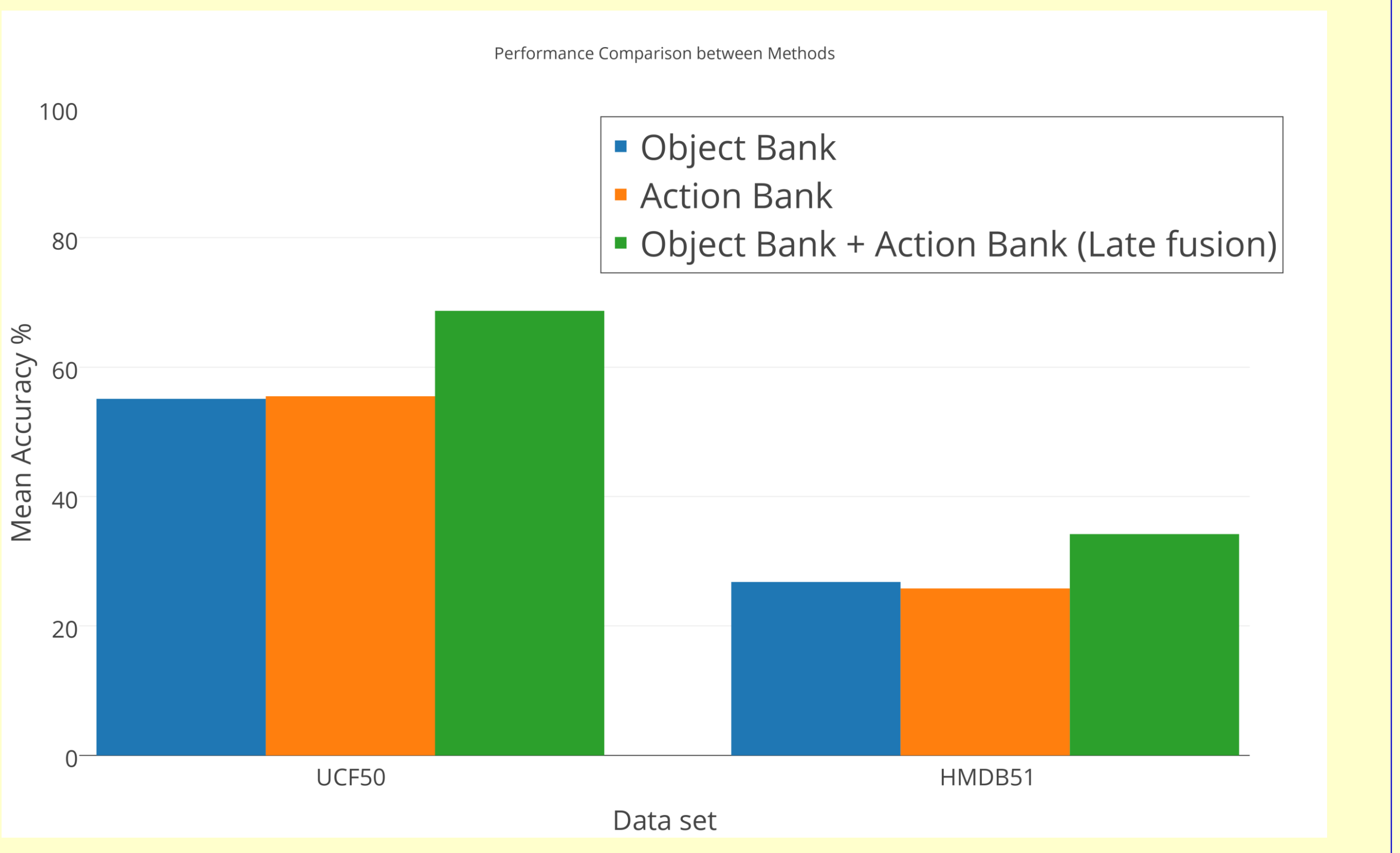
◆UCF50 & HMDB51 Data set:
Over 6000 videos which more than 50 categories in each datasets

◆Cross-validation: Leave-one group out & Three train splits

◆Evaluation Method: Macc.
Mean Accuracy

Example: Key frames from the UCF50 video data set

Classification results



Comparison with each method individually, And with late fusion of OB and AB

6. Conclusion

- Object Bank & Action Bank are complementary when performing fusion on these features
- Promising potential with improvements in the quality and number of action and object detectors
- Future work including detailed investigation on deep learning methods