BUPT-MCPRL@TRECVID 2016: Multimedia Event Detection

Rui Xiang,
Zhicheng Zhao, Fei Su, Shizhe He

Multimedia Communication and Pattern Recognition Labs,
Beijing University of Posts and Telecommunications
(**BUPT-MCPRL**)

hgjngh123@bupt.edu.cn






Chapter Structure

- MED Review

- Framework Introduction
  - introduction of methods
  - experiments result
  - conclusions and discussions

- Our Results

MED Review

Our results **last year** :
Testing on MED15-PS-EvalSub (using infAP200)

Task	10EX	100EX
our result	0.087	0.155
other teams' best result	0.303	0.365

- It's the 1st time we take part in the MED task.
- The Events are complex, so the method should be robust enough.
- The method should not rely on large scale training samples

Our performance **this year** :

Testing on MED16 Pre-Specified Events (using MAP)

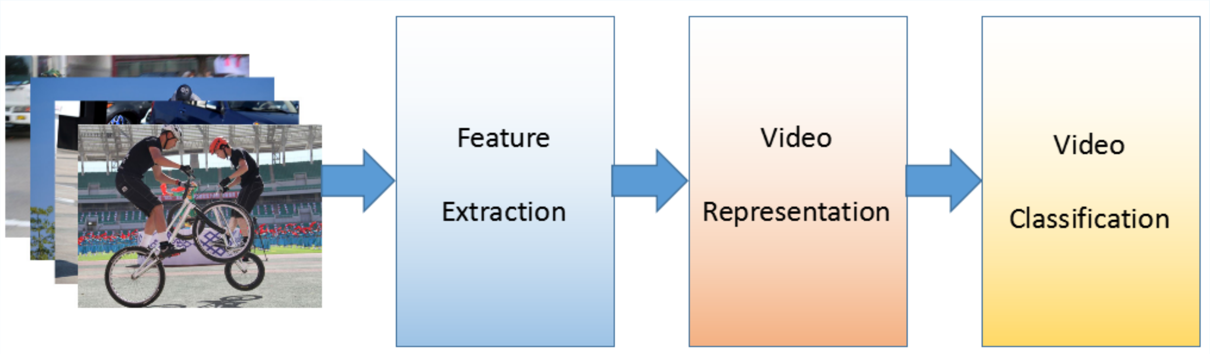
Teams	PS_SUB_10EX	PS_SUB_100EX	Platform
Our p-baseline	**0.336**	**0.469**	**SML**
Our c-contrast(Progress)	**0.354**	**0.490**	**SML**
Etter	0.014		SML
INF	0.298		SML
ITICERTH	0.318	**0.462**	SML
KU-ISPL	0.209	0.340	SML
MCIS	0.004	0.004	SML
MediaMill(FullAsSub)	**0.354**		SML
NIIHitachiUIT	0.007		SML
TokyoTech	0.279	0.415	SML
VIREO	0.335	0.419	MED
nttfudan	0.328	0.457	SML

Framework Introduction

key words:

- high performance, high speed, low storage cost

structure of our framework **:**



strategy:

- p-baseline:choose the best method for each module.
- c-contrast:fuse most effective methods.

Feature Extraction

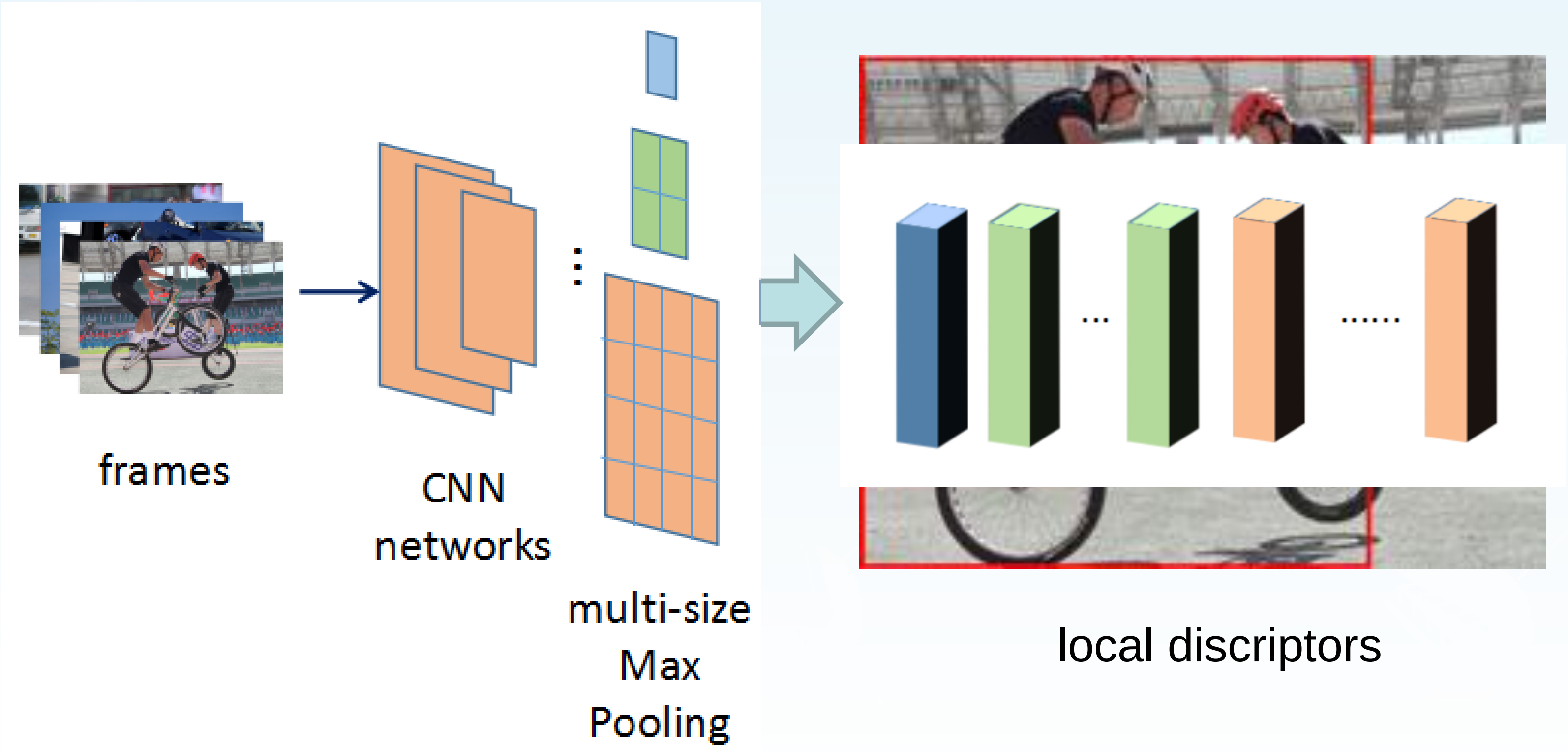
3 Strategies of CNN feature extraction:

- Global Descriptor
  extracted from the fc7 or last average pooling layer.

- Dense Local Descriptor
  extracted from multi-size pooling layers

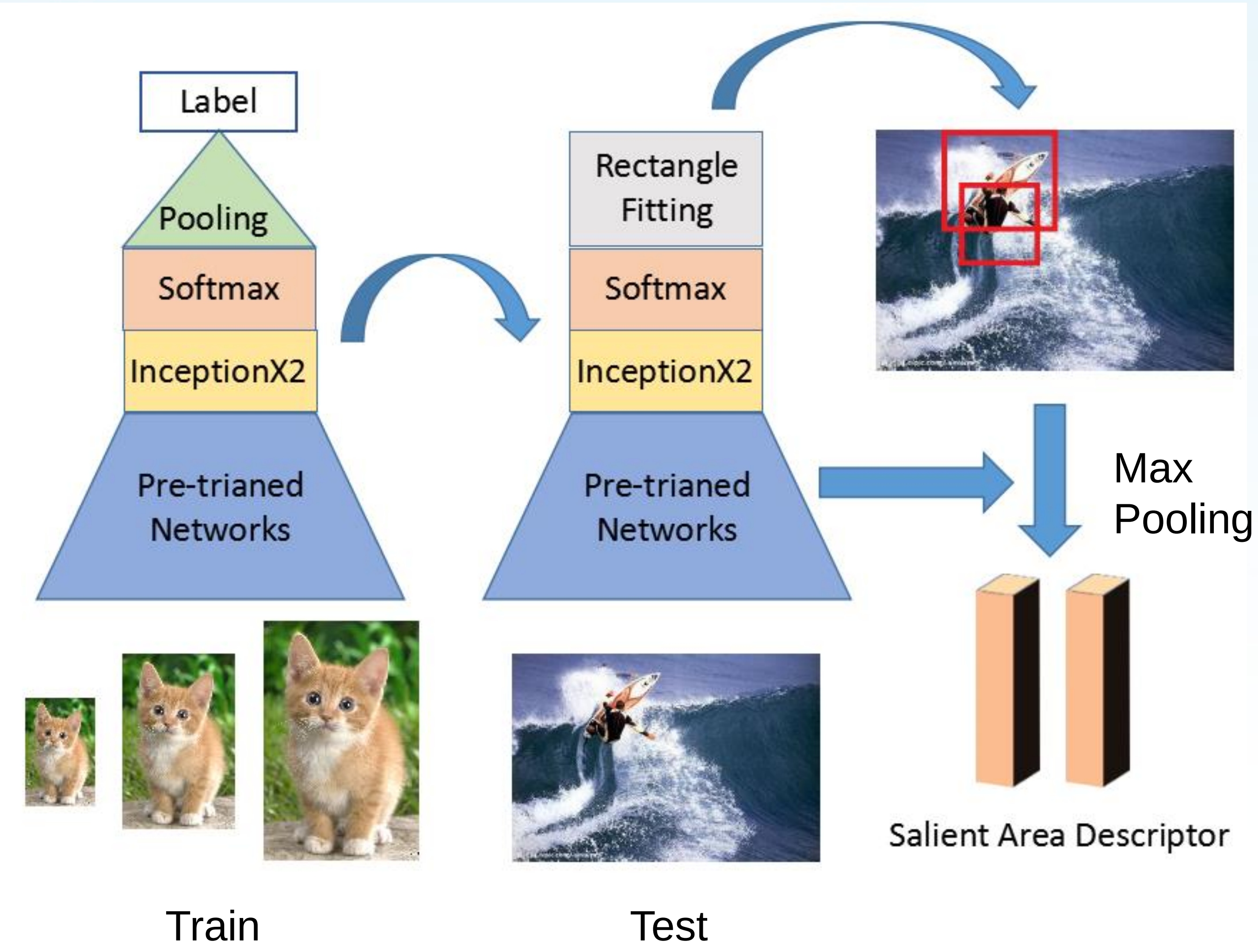
- Salient Area Descriptor
  extracted from a fast Region Proposal Network.

Dense Local Descriptor



The local descriptors provide different areas' information like a sliding window moving on the image.

Salient Area Descriptor



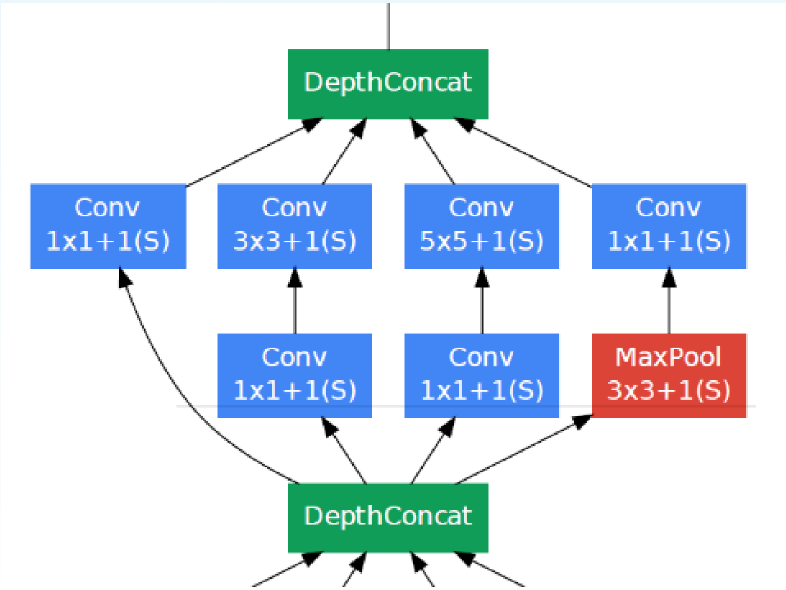
Experiment on strategies of feature extraction

method	Storage Cost	Precision(MAP)
Global	13G	0.481
Dense	367G	**0.496**
Salient	46G	0.495

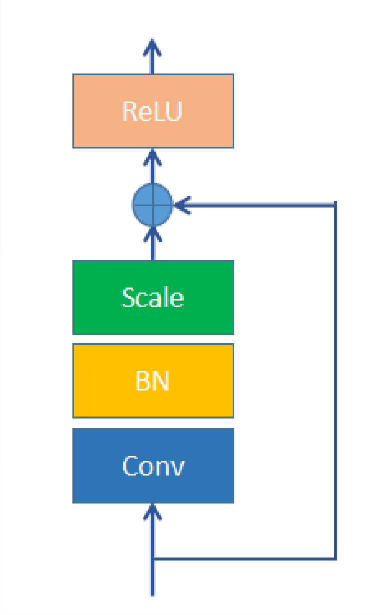
train: 800 videos
test: 7230 videos

- We choose the dense local descriptors as our baseline methods for its best performance and simple form.
- Salient method is also competitive for its low storage cost without significant drop of result.

CNN Model Selection:



Inception



residual

- The inception structure provides ability of fast compute
- The residual network can be very deep and more precise

Video Representation and Classification

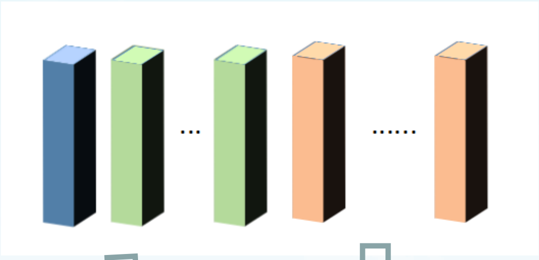
Baseline Methods:

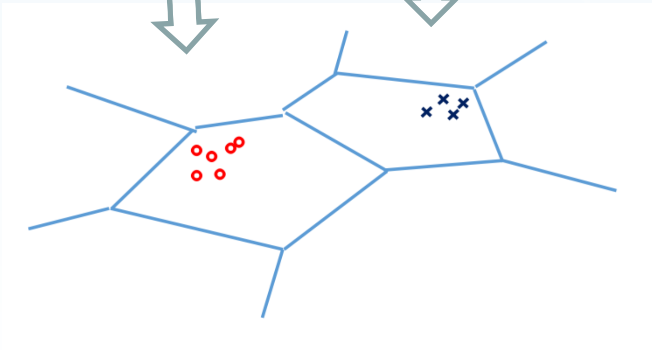
- VLAD for video representation
- SVM for video classification

Extending Methods:

- FisherVector
- Rank SVM
- LSTM
- netVLAD
- temporal kernel CNN

Video Representation





VLAD encode
The residuals of samples to nearest clusters are storaged.

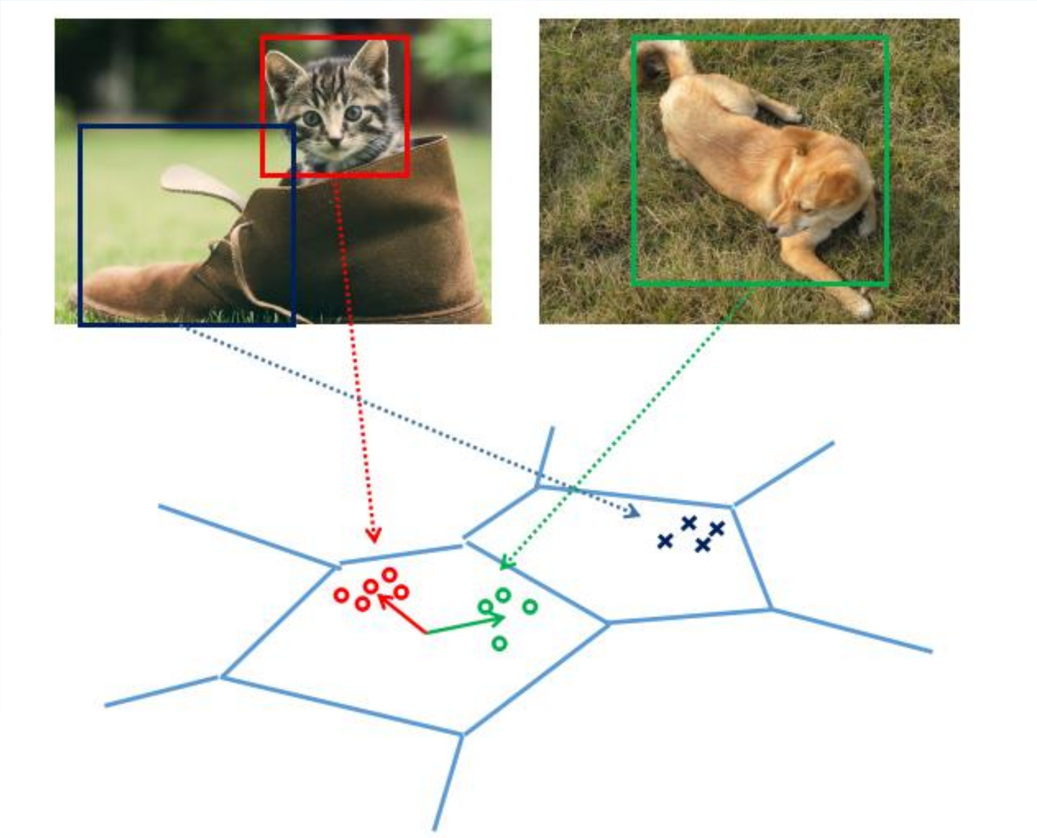
k-nearest assignment VLAD :

$$v_{i,j} = \frac{1}{N}\sum_{n=1}^{N} \alpha_i (x_j^{(n)} - c_{i,j})$$

$$\alpha_i = \begin{cases} \dfrac{\exp\left(-\gamma dist(x,c_i)\right)}{\sum_{j=1}^{K} \exp\left(-\gamma dist(x,c_j)\right)} & , dist(x,c_i) \in topk\{dist(x,c_j)\} \\ 0 & , dist(x,c_i) \notin topk\{dist(x,c_j)\} \end{cases}$$

$$dist(x,c_i) = \sum_{j}^{D} \left(x_j - c_{i,j}\right)^2$$

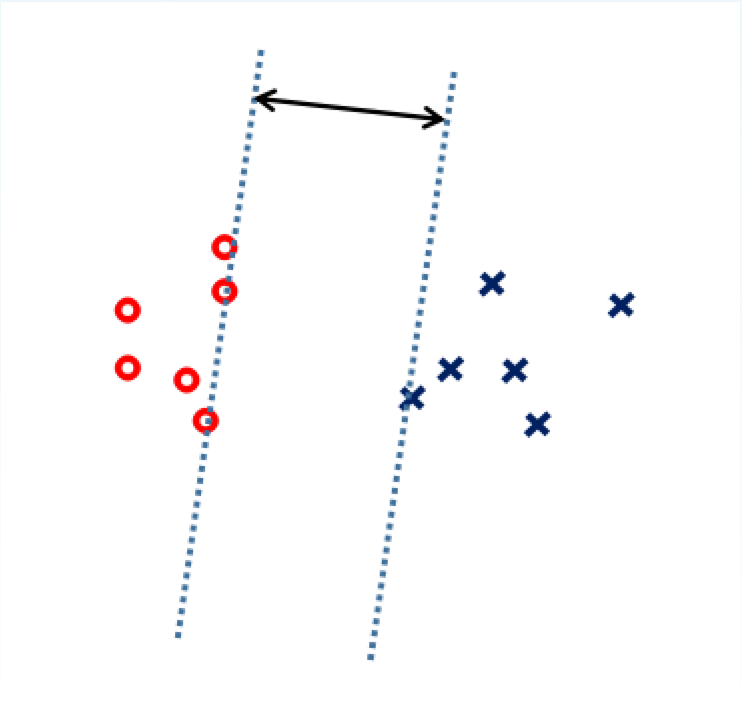
$v_{i,j}$ is the value of i-th cluster and j-th dimension



The explanation of the VLAD

Every attribute of a video will be comapred when we calculate the distance.

Video Classification



Support Vector Machine

Linear SVM for binray classification:

$$\begin{aligned} &Minimize \quad \frac{1}{2}\|w\|^2 \\ &s.t. \quad y_i(wx_i + b) > 1 \\ &i = 1,2,\ldots,N \end{aligned}$$

Score prediction:

$$score = \frac{1}{1 + \exp\left(-\alpha(wx + b)\right)}$$

$$\alpha \in (0, +\infty)$$

Experiments

- Comparision between different video representations:

Method Name	Result(MAP
VLAD	**0.232**
FisherVector	0.228
Rank SVM	<0.1
C3D(without re-training)	<0.1

dataset: MED14_Progress
model: VGG16
train: 5030 videos
test: about 30000 videos

- Comparision between Deep learning methods:

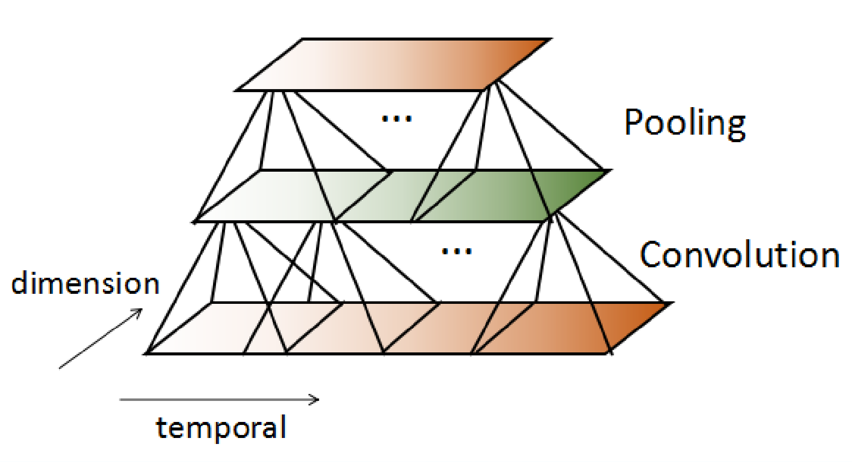
Method Name	Result(MAP)
VLAD+SVM (p-baseline)	**0.640**
LSTM	0.382
netVLAD	0.525
Temporal convolution	**0.565**

dataset: MED16_TRAIN
model: GoogleNet12988c
train: 7230 videos
test: 800 videos

- Our p-baseline method outperfoms others by a large margin

discussions

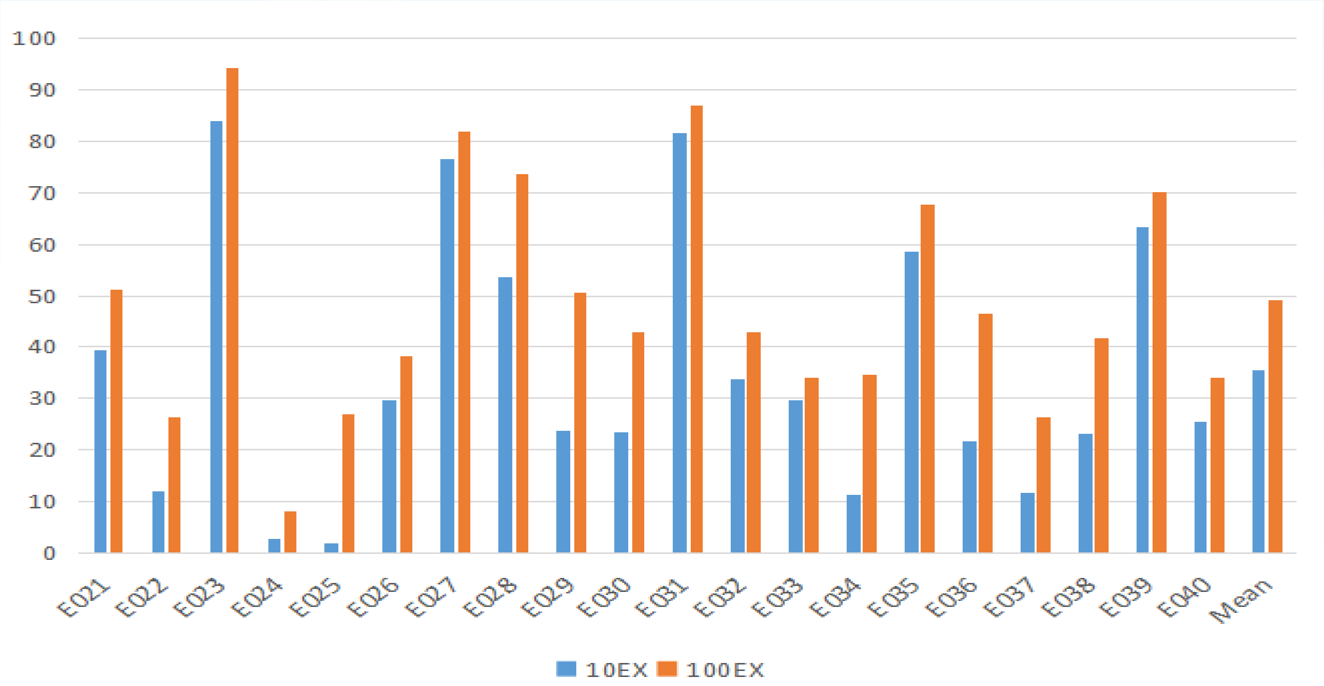
- The Temporal Convolution structure is very potential.



- Deep learning methods may benefit from a end to end video classification modle. This can be realized by combining Spatial Convolution and Temporal Convolition.

Our Results

Result(MAP)	PS_SUB_10EX	PS_SUB_100EX	Platform
Our p-baseline	0.336	0.469	SML
Our c-contrast(Progress)	0.354	0.490	SML



Thank You !

http://mmc.sice.bupt.cn

BUPT-MCPRL