A Finer Granularity Object-level Representation for Multimedia Event Detection

Speaker: Hao Zhang

Hao, Zhang, Yi-jie Lu, Chong-Wah Ngo
City University of Hong Kong

TRECVID 2016 Workshop, Gaithersburg, MD, USA

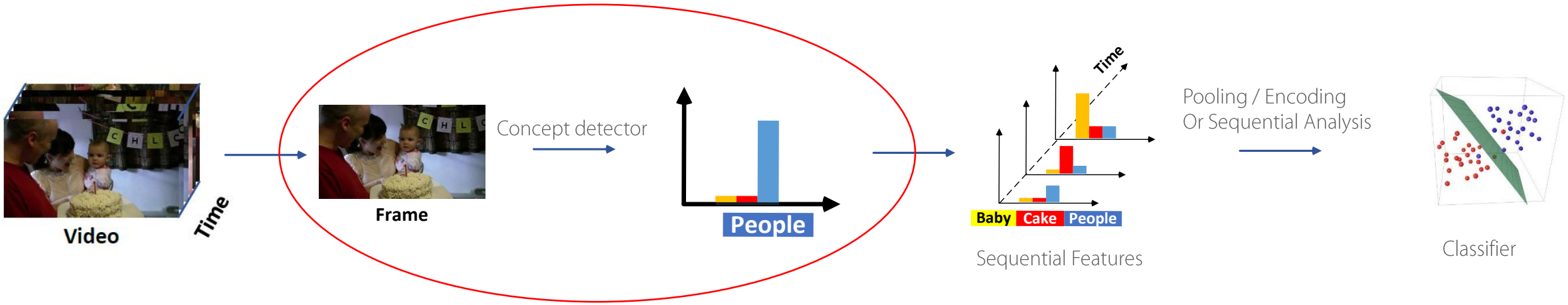
VIREO Video Retrieval Group

MED: Semantical Video Representation.

Event Detection

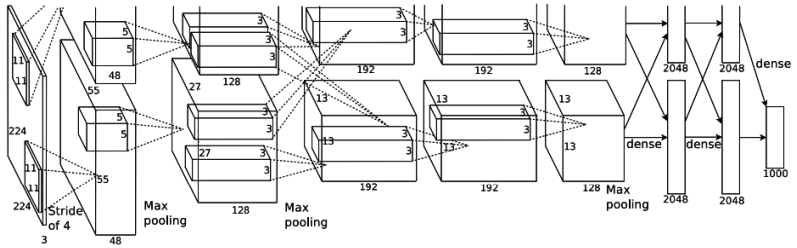
semantical relation

Video → Event Topic

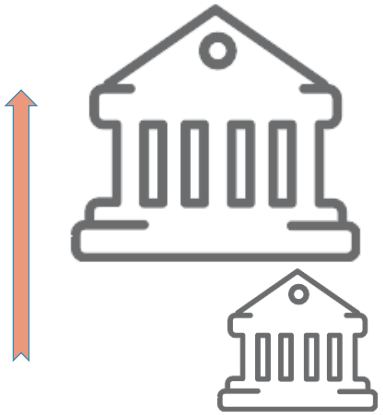


Motivation

General Representation for Video → Concept Bank



DCNN based concept detector on single-labeled dataset



ImageNetShuffle-13K
ImageNetShuffle-8K
ImageNetShuffle-4K

ImageNet-1000

Scale Up

Places-205
SIN-346
Research Collection-497
FCVID-239
Sport-487

(number denote vocabulary size)

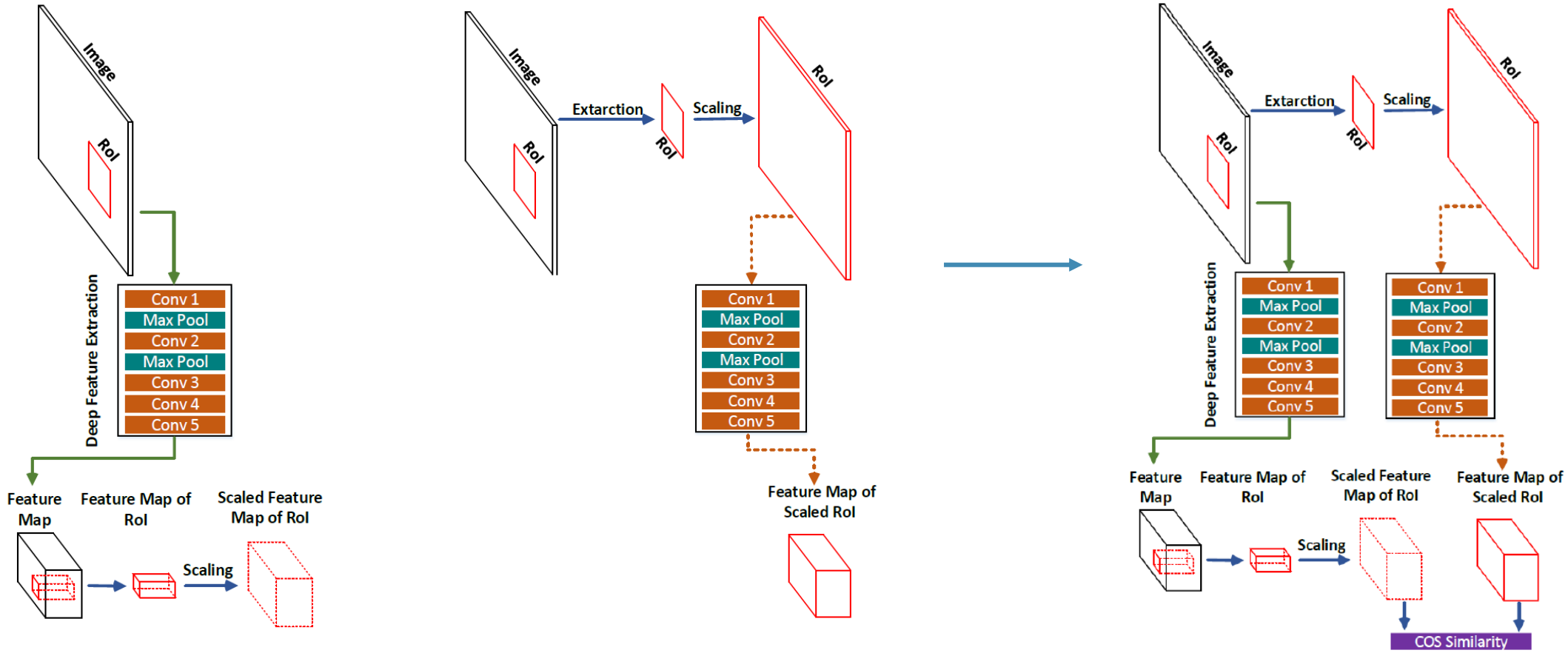
A very sparse semantical vector for frame:

Vocabulary larger **Vector sparser** →

Emphasize primary object
Overlook small size regional object

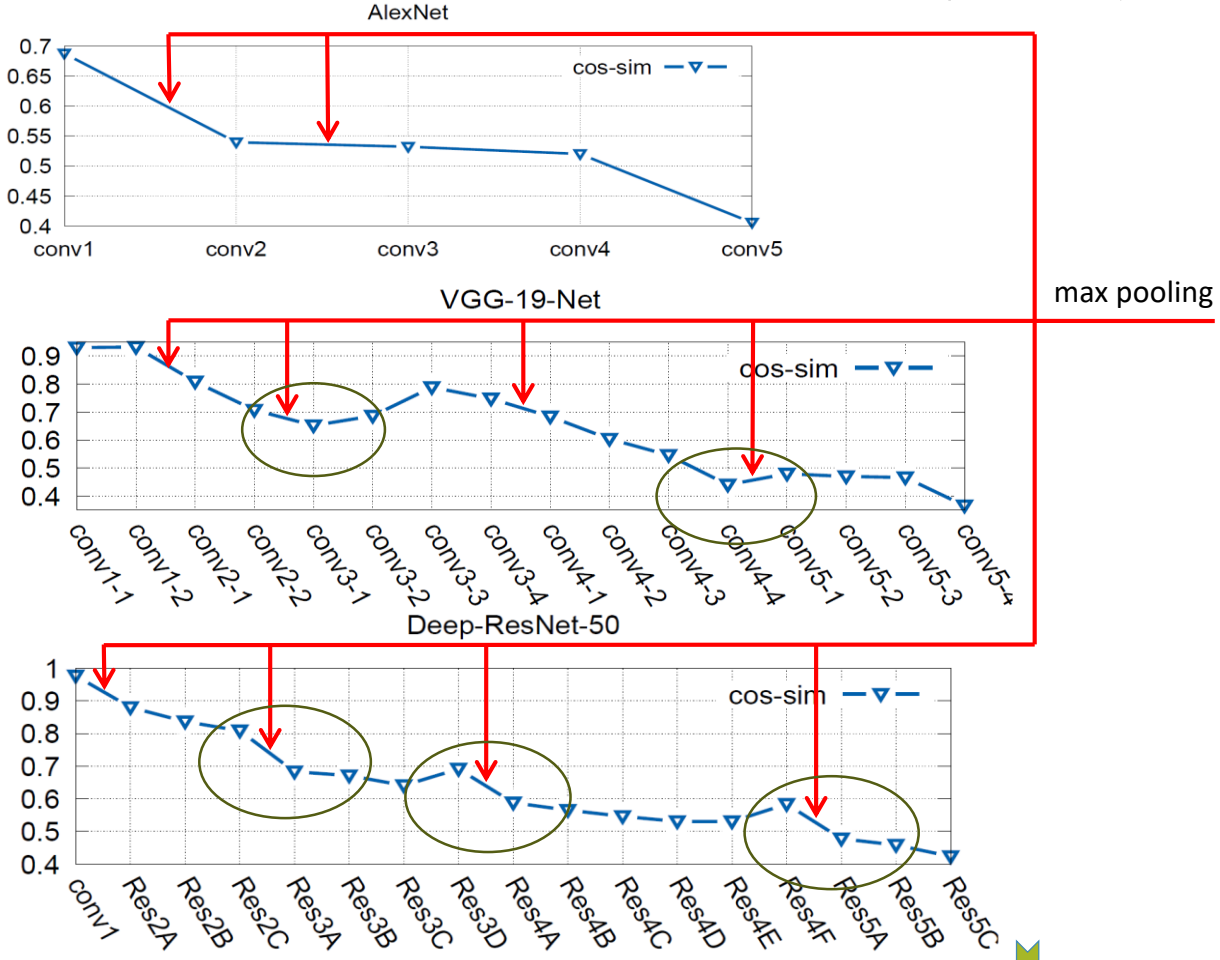
Motivation

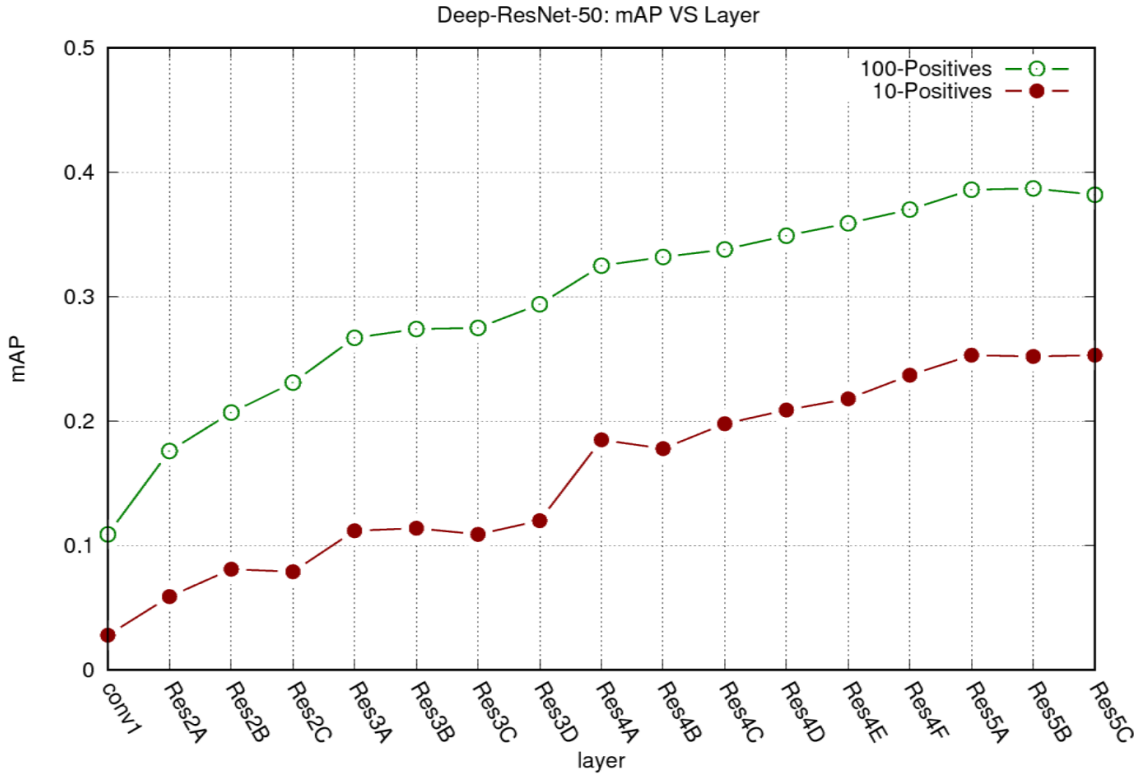
Measurement of regional information loss



Motivation

Measurement of regional information loss (AlexNet, VGG-19-Net, RestNet-50)





Adopt CNN-VLAD with different deep features for MED

Only around **40%** of Regional Information left

Discriminatory power of deep features consistently **improves**

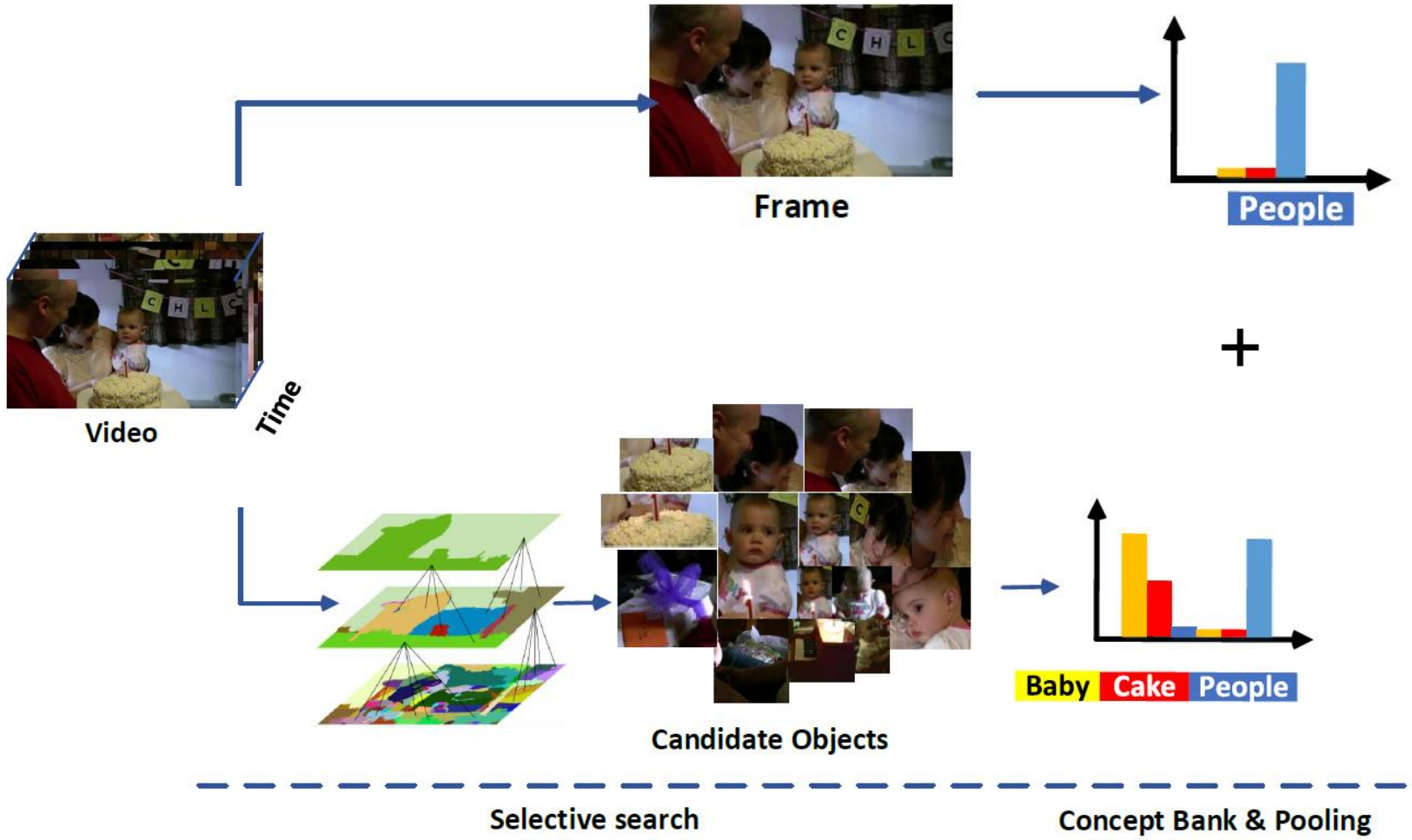
Verifications:

***-Framework-1**:*
Object-Pooling: Apply Concept Detectors on RoI

***-Framework-2**:*
Object-VLAD: Encoding RoI

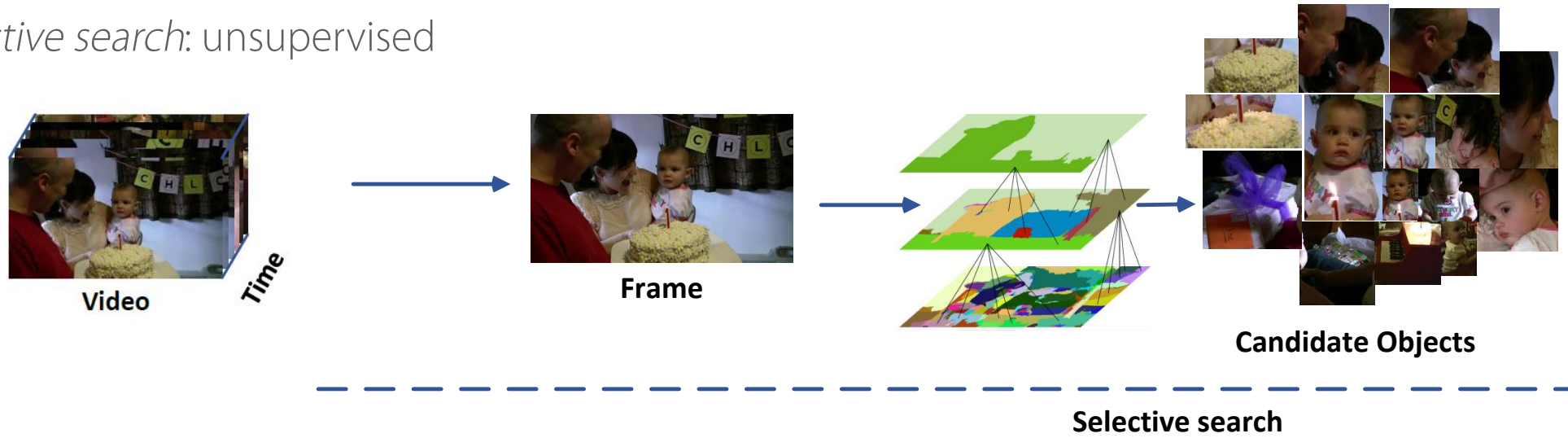
Framework-1: Pooling RoI

Object-Pooling: Apply Concept Detectors on RoI



Framework-1: Candidate Object Proposal

Selective search: unsupervised



Candidate Object Proposal:

1. Selective Search *(single strategy)* proposes around 200 alternative regions.
2. **Remove tiny** regions (areas < 1/10 of frame)
3. **Remove bar shape** regions (height/width or width/height ratio > 4)
4. Retain only one, if two regions has large overlap (intersection/ Union > 0.5)

On average, each frame has 20 candidate object regions.

Framework-1: Selective Search

- **Observations**:

selective search segments objects better than objectness, BING, edgebox:

- Possible reasons:

1. Initially designed for Images.
2. Video frame suffer from motion blur
3. Color and brightness > saliency, edged based.

-Alternative method:

Selective search vs ImageNet-200 Objects:

1. ImageNet-200 objects is supervised, but the **vocabulary size is small** for varieties of multimedia events.
2. Selective search is unsupervised and is generalizable to new events.

Framework-1: Experiments

With different concept bank :
SIN-346, Research Collection-497, ImageNet-1000, Concept-Bank (Combines all)

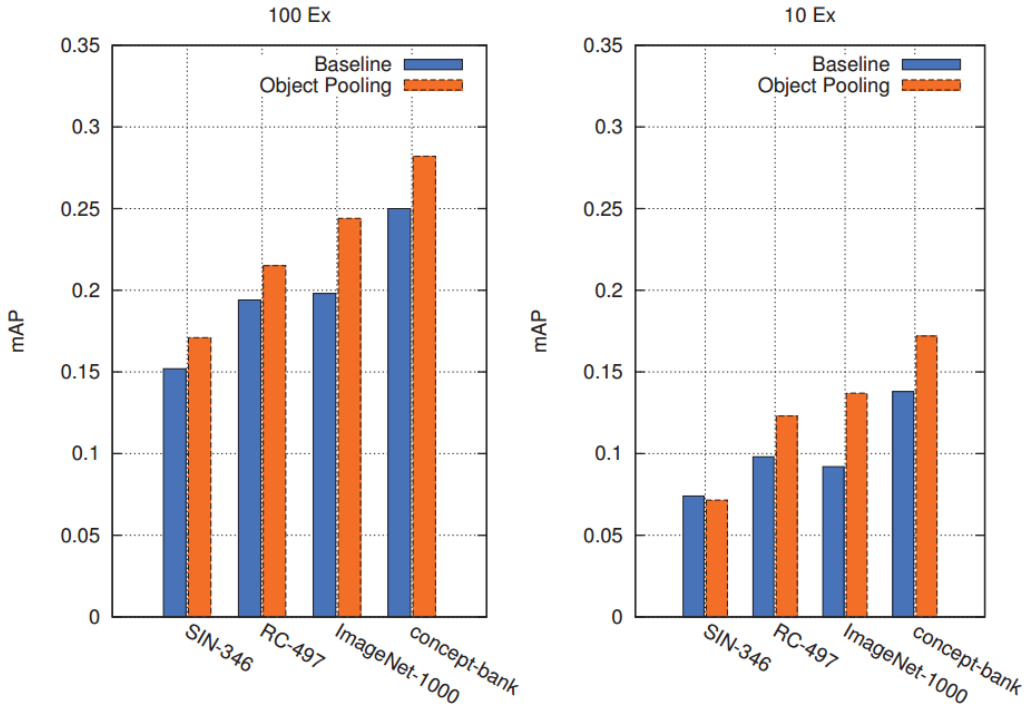


Fig. 6 MED14_Test 100/10Ex performance comparision with different concept sets (mAP)

1. Object-level pooling is complemented to Baseline, and brings improvement.
2. New representation is robust to different concept sets.
3. Vocabulary grows larger, improvements become larger.

Framework-1: Experiments

Replace concept feature with deep features from middle layers

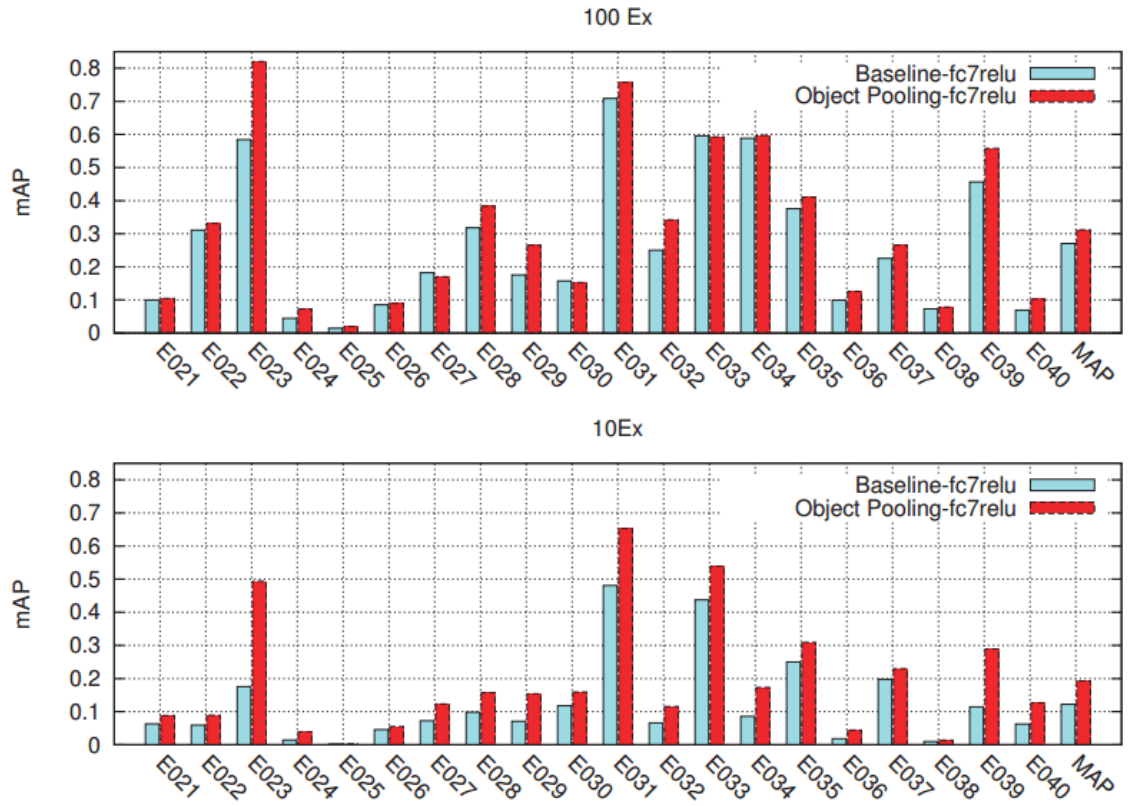
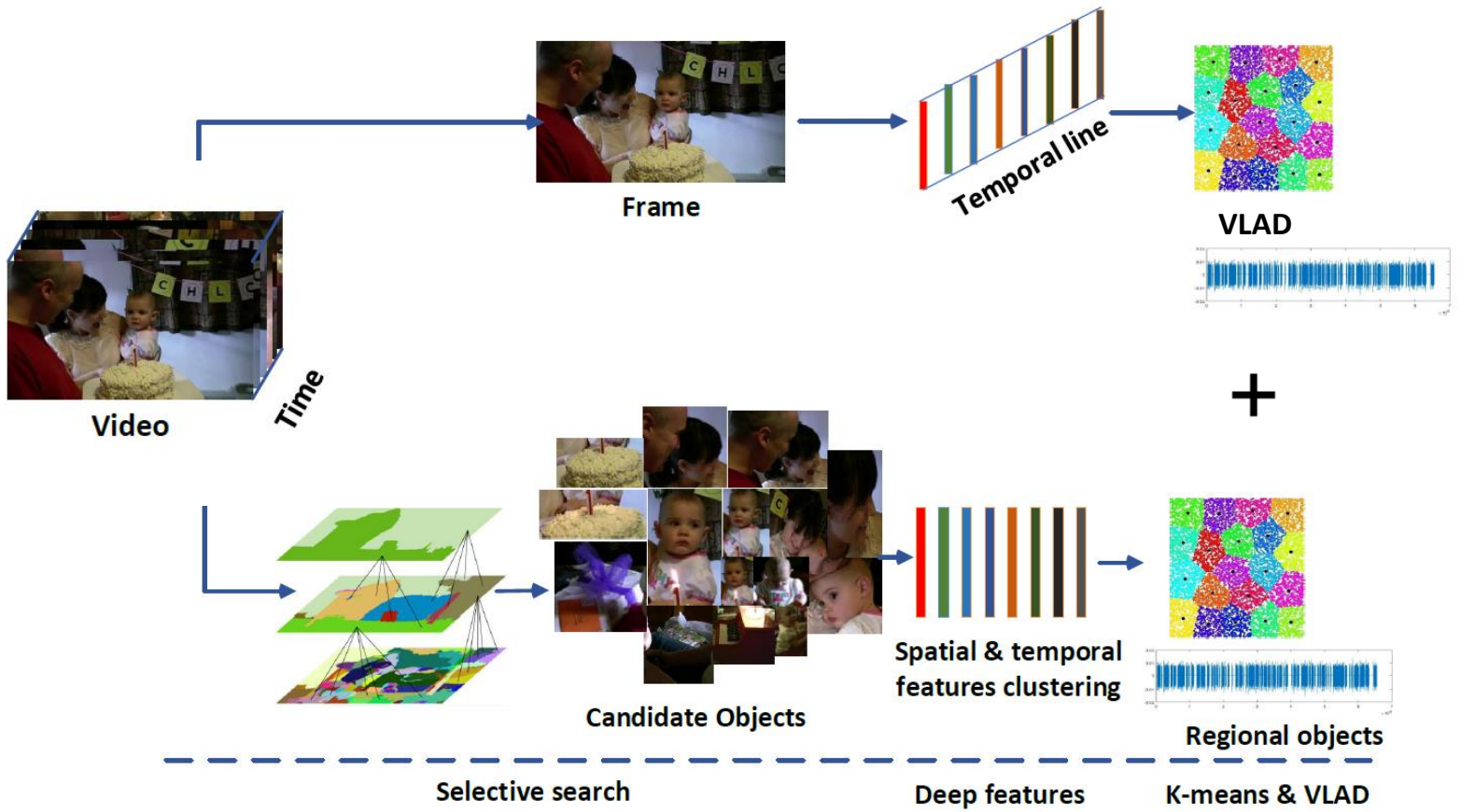


Fig. 5 MED14_Test 100/10Ex per event performance comparision with *latent-concept* descriptors (mAP in percentage)

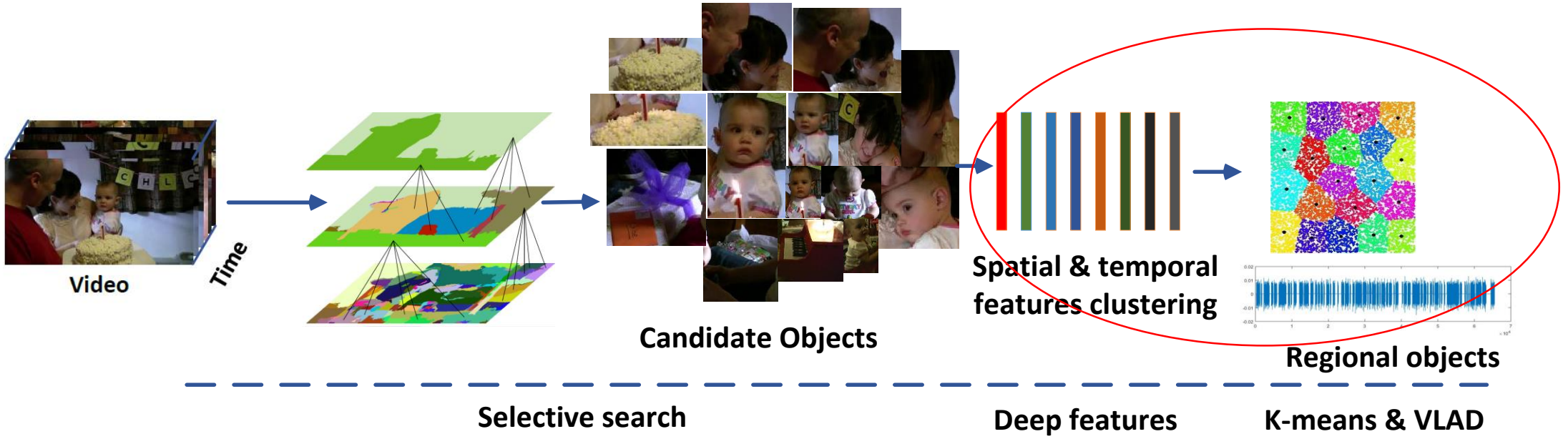
Framework-2: Encoding RoI

Object-VLAD: Separately encoding global and regional information



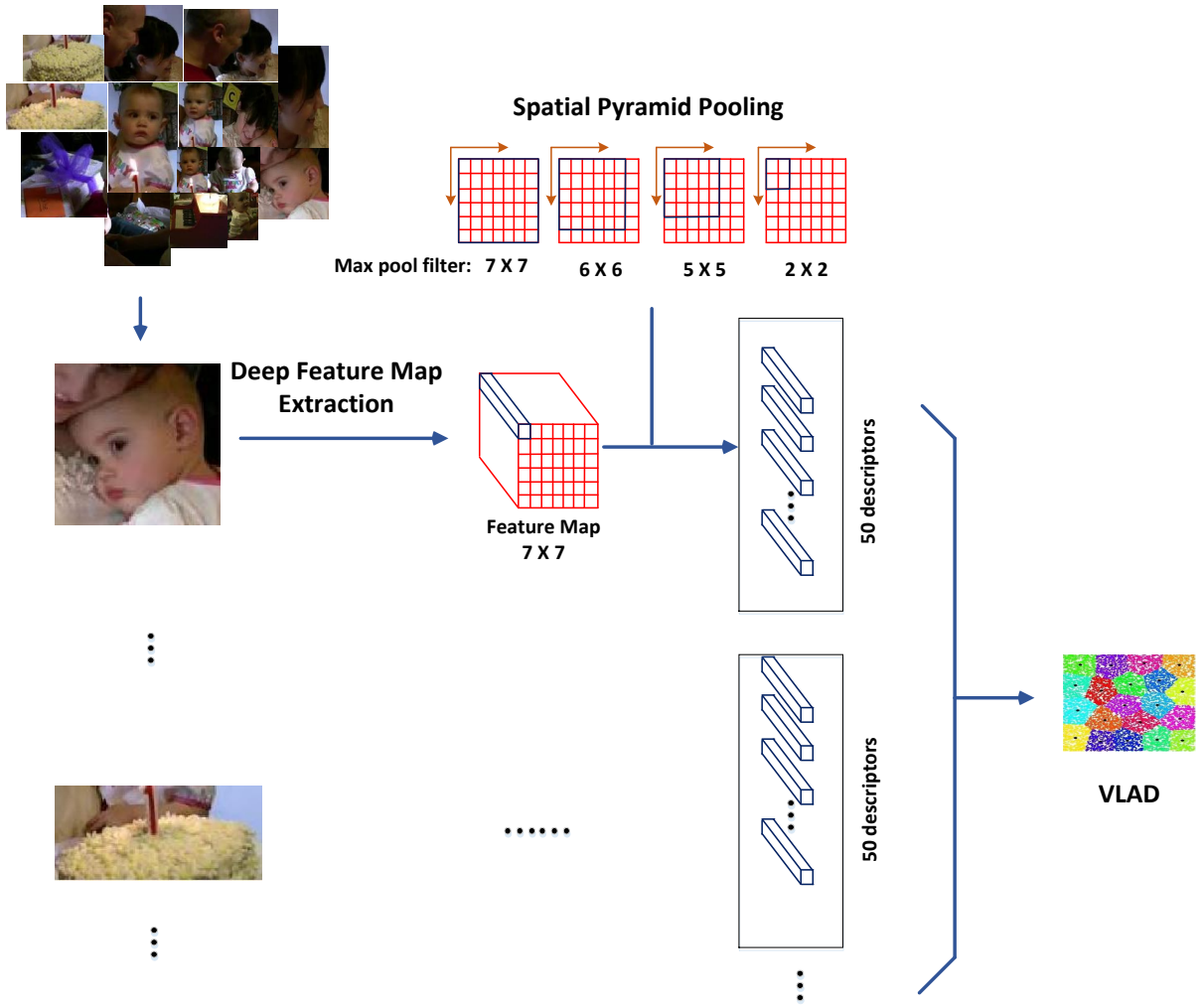
Framework-2: Encoding RoI

VLAD Details



A Discriminative CNN Video Representation for Event Detection
Zhongwen Xu, Yi Yang, Alexander G. Hauptmann (CVPR'15)

Framework-2: Spatial Pyramid Pooling Details



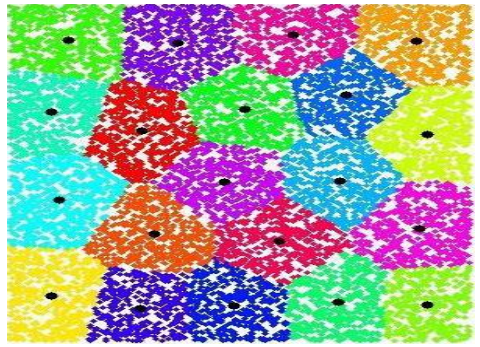
- ***Feature***:
res5A feature map from *ResNet-50*

- ***Spatial Pyramid Pooling filter:***
7 X 7, *6 X6*, *5 X 5*, *2 X 2*

Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (ECCV'14)

Framework-2: VLAD Details

VLAD



$$\boldsymbol{l}_k = \sum_{i=1}^{I} \sum_{n=1}^{N} q_{ink} (\hat{\boldsymbol{y}}_{in} - \boldsymbol{\rho}_k)$$

i-th frame

n-th object

k-th cluster center

$$\Phi(\hat{\boldsymbol{Y}}) = [\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_k, \ldots, \boldsymbol{l}_K]$$

- PCA with whitening
  dim of descriptor 2048 -> 256 (with 99% energy retained)
- K-means (Vocabulary Size)
  256/128 (VLFeat K-means++)
- Soft-assignment (nearest 5 cluster center)
  q = 1/5
- Other VLAD options (VLFeat)
  square root & normalize components

Framework-2: Experiments

CNN-VLAD vs *Object-VLAD*

CNN-VLAD

Encoding Frame-level features.

Object-VLAD

Selective search proposes candidate objects.

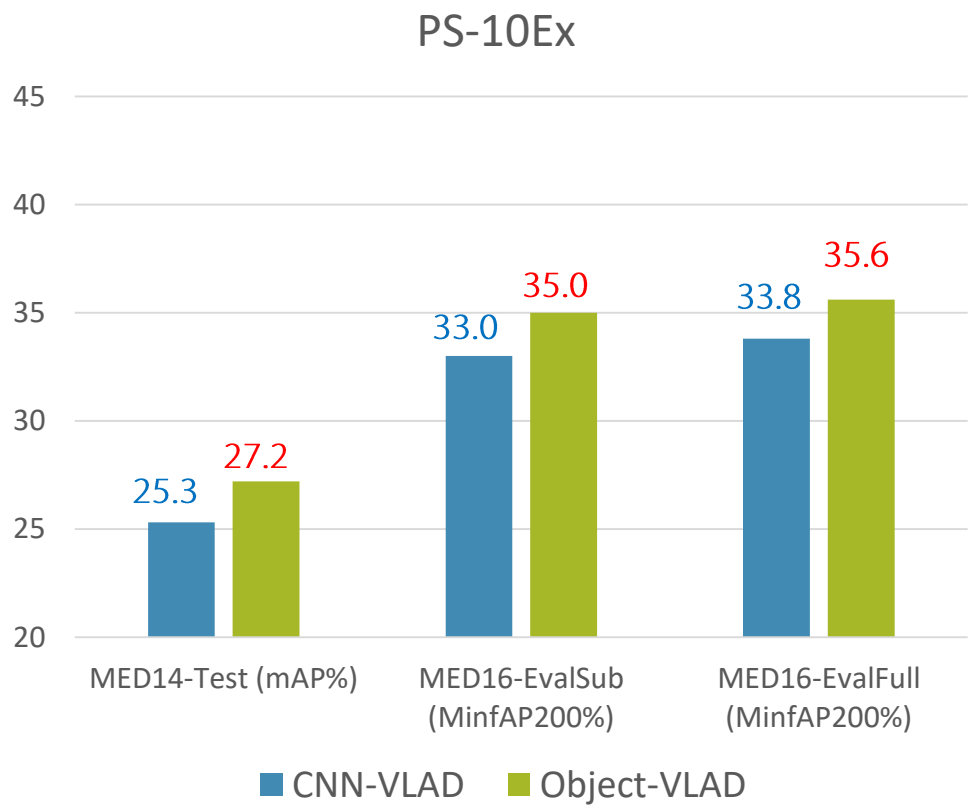
Encoding Frame-level features and Objects

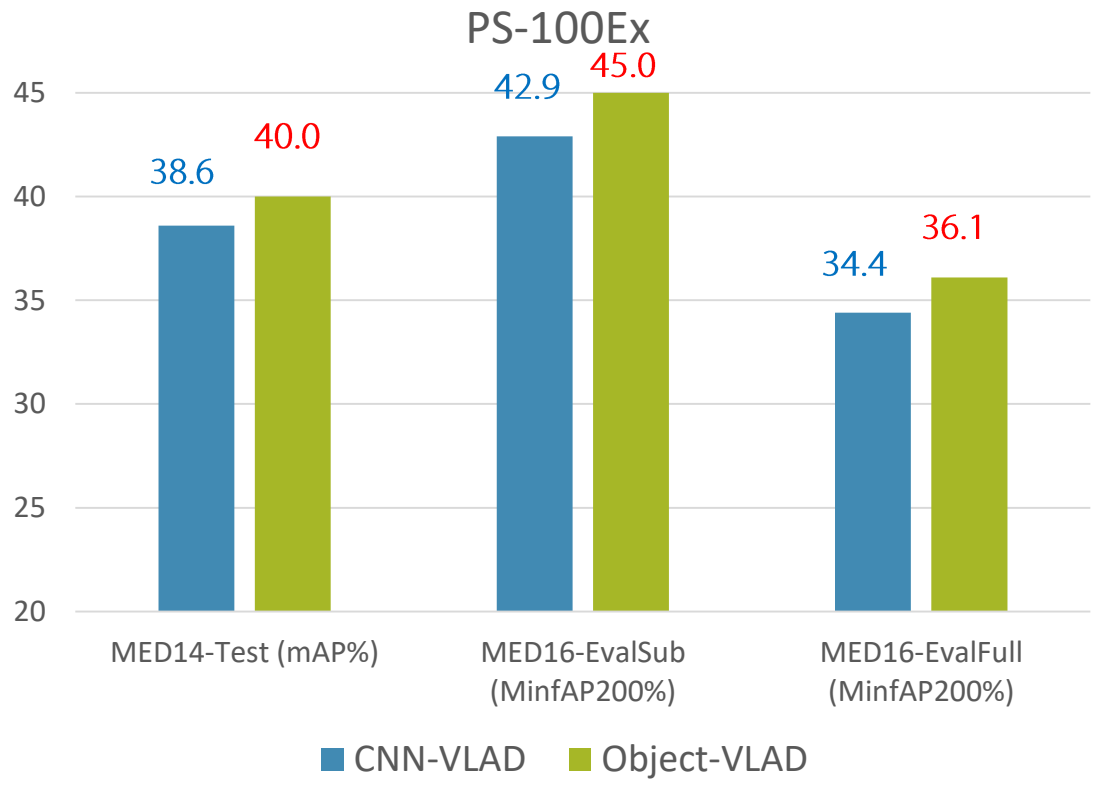
res5A feature map (ResNet-50): VLAD Encoding 2048 Dim latent descriptors (obtained by SPP [2])

Linear SVM

Framework-2: Experiments

Regional information (res5A feature from ResNet-50)





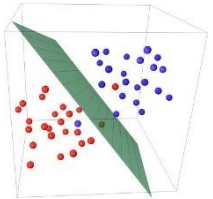
Framework-2: Experiments

Concept-Bank vs *Object-VLAD*

Concept-Bank

ImageNetShuffle-13K
ImageNet-1000
Places-205
SIN-346
Research Collection-497
FCVID-239
Sport-487

15,762 Concepts
(Most concept detectors are fine-tuned with Deep RestNet-50 architecture)



Chi2 SVM

Object-VLAD

Selective search proposes candidate objects.

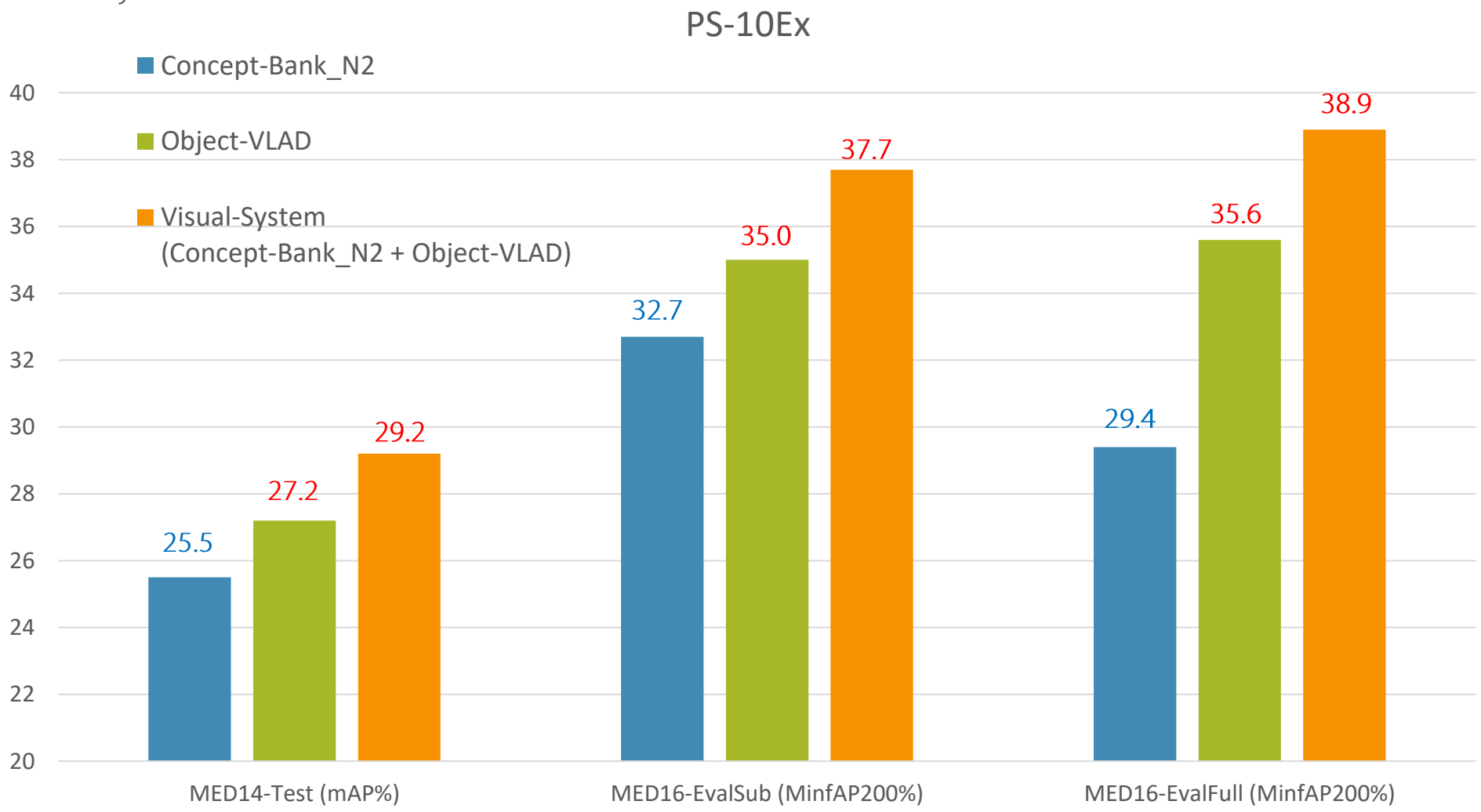
Describe objects by res5A feature from Deep ResNet-50 (trained on ImageNet-1000)

VLAD Encoding 2048 Dim latent descriptors (obtained by SPP [2])

Linear SVM

Framework-2: Experiments

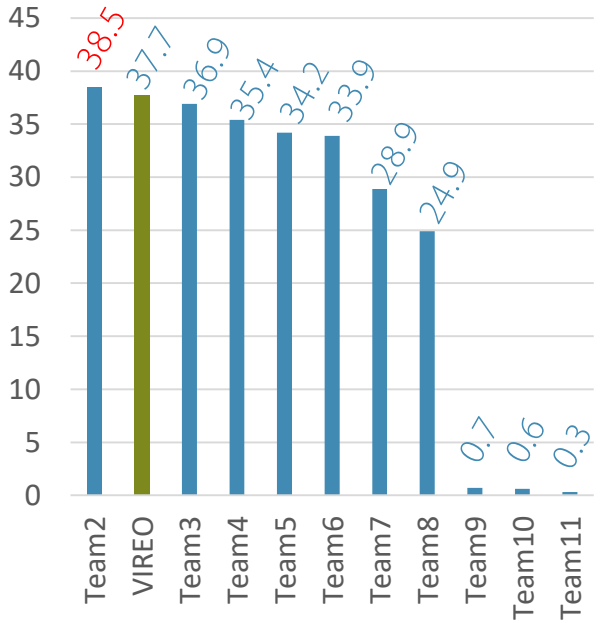
Concept-Bank VS Object-VLAD

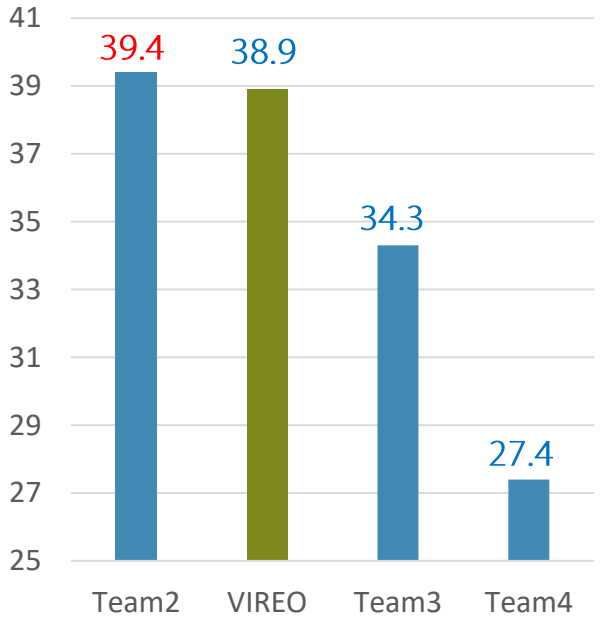


TRECVID-2016: Primary Evaluation

PS-10Ex & AH-10Ex

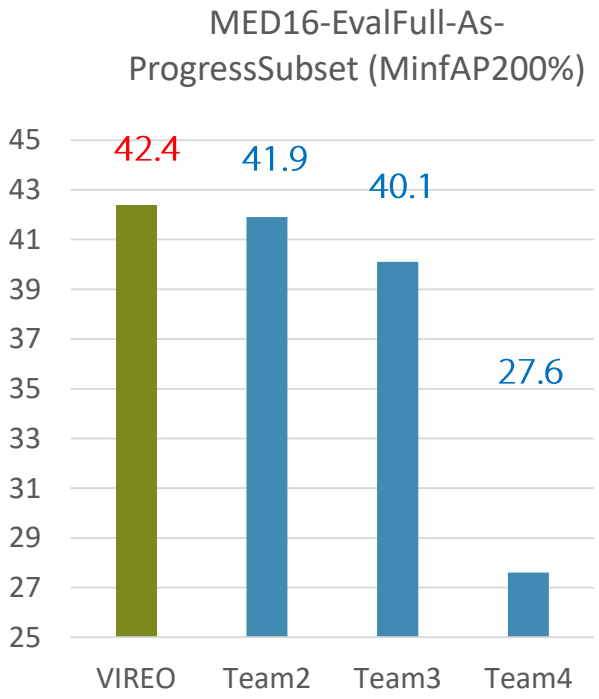
Object-VLAD (Framework-2) + ***Concept-Bank***

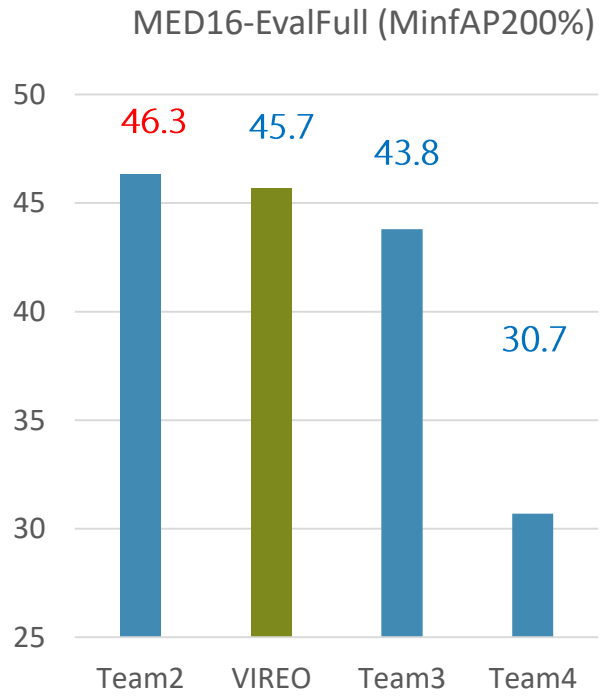




AH-10Ex

Object-VLAD (Framework-2) + ***Concept-Bank***



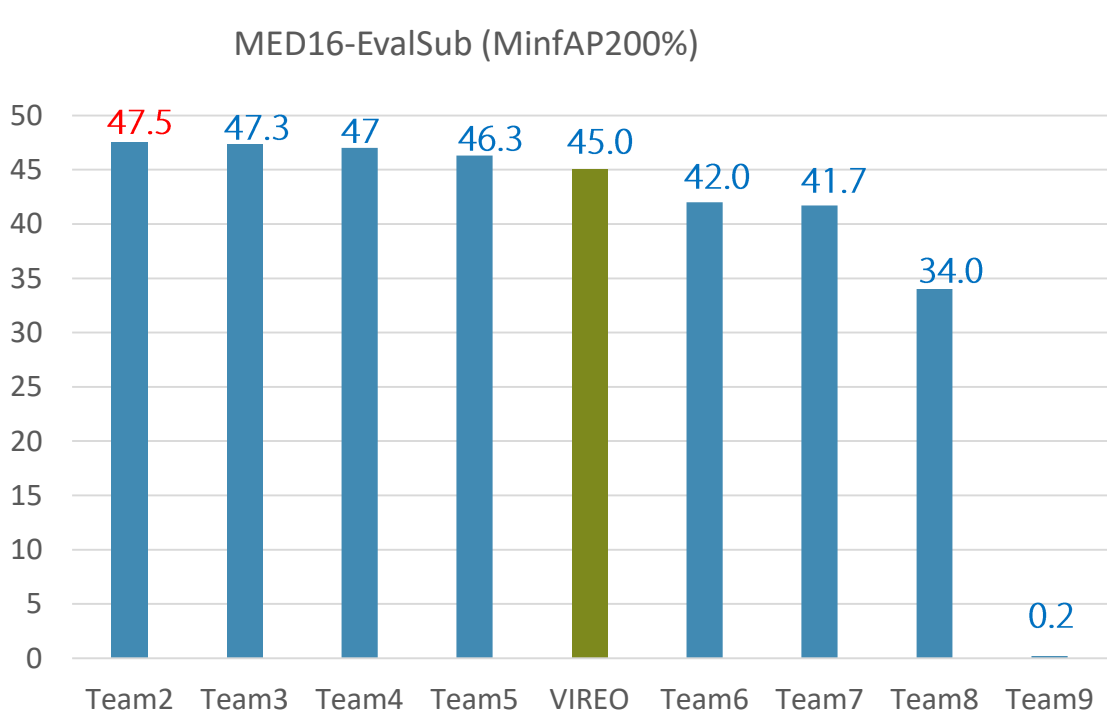


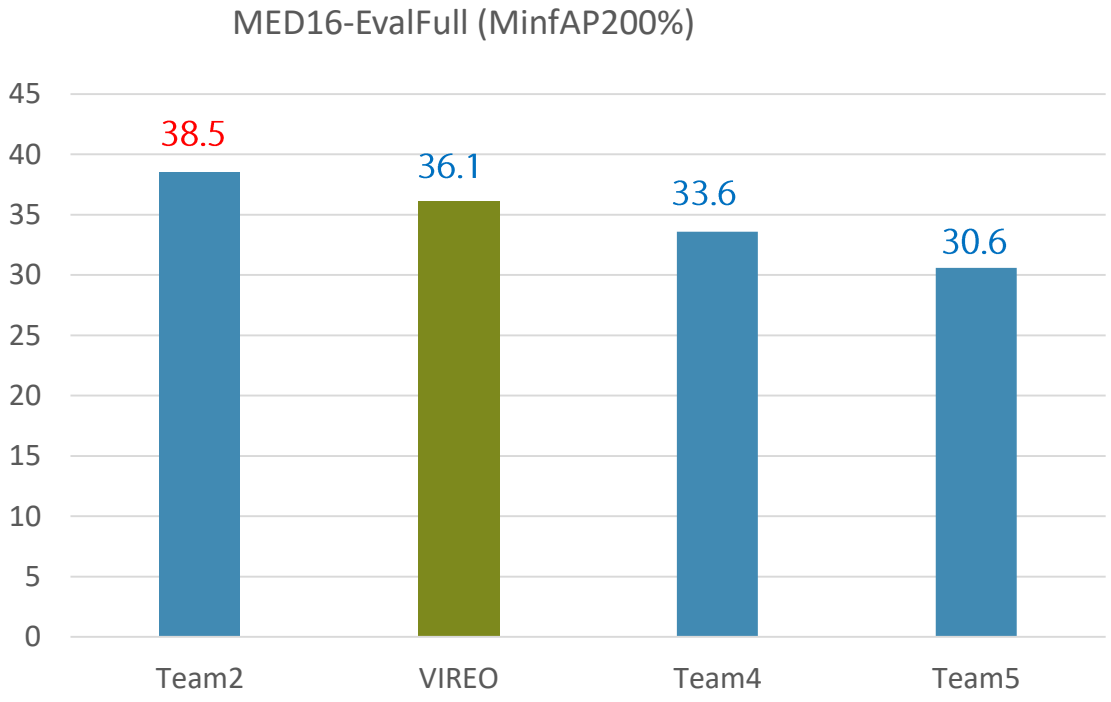
*Combing only semantical and object-level features, our model achieved **top-2** performance in PS/AH-10Ex subtaskes.*

TRECVID-2016: Primary Evaluation

PS-100Ex

Object-VLAD (Framework-2)





*With **only** object-level video representation, our model achieved acceptable performance in PS-100Ex subtask.*

Thanks & QA