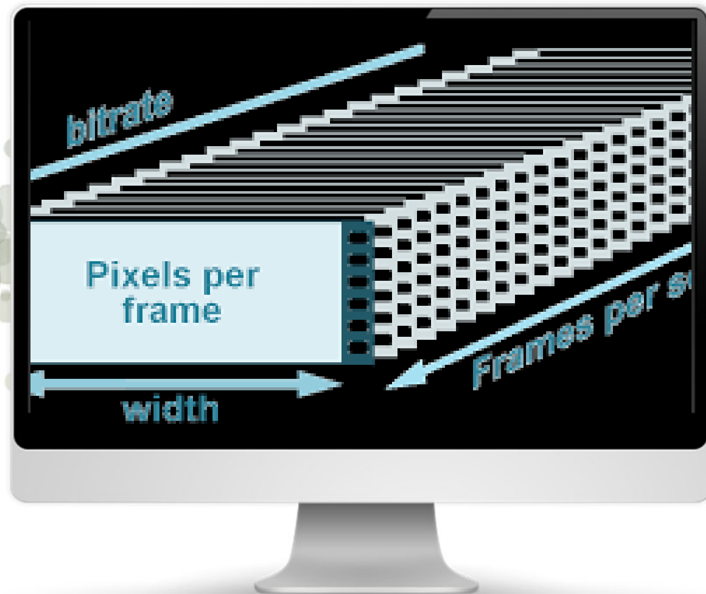


VIDEO REPRESENTATION

Supervisor: Dr. A. Mansouri

Amir Hossein Bakhtiari

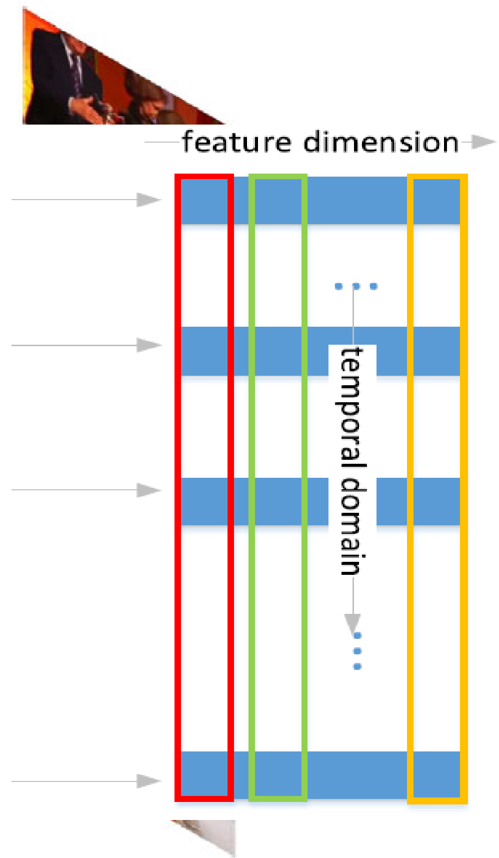
What is video representation?

- A video is composed of a sequence of frames, which reflects the evolution of the video content.
- A video can be *represented* by a sequence of frame-level features

First Person Video (Egocentric)

The main difference between conventional 3rd-person videos and 1st person videos is that, in 1st-person videos, the person wearing the camera is actively involved in the events being recorded.

Frame-level Features (Descriptors)



Types of Frame-level Features

- **Appearance Features:**
  CNN features of each video frame

- **Motion Features:**
  Histogram of Oriented Gradients (HOG)
  Histogram of Optical Flows (HOF)
  Improved Dense Trajectory (IDT)

Pooled Motion Features for First-Person Videos

01 Per-frame feature representation

02 Time series representation

03 Temporal pooling

04 Final representation

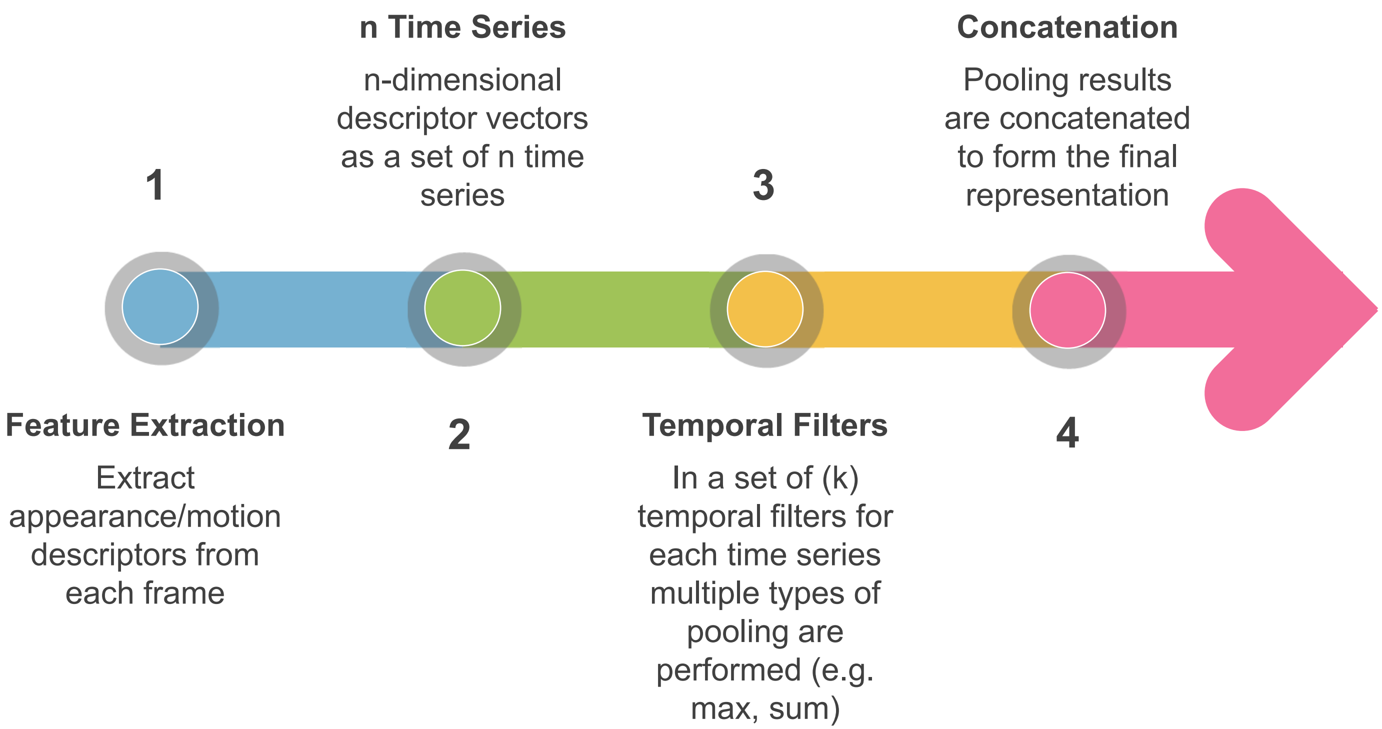
Pooled Motion Features for First-Person Videos

Figure 2. Overall representation framework of our pooled time series (PoT).

Pooling: pooled time series (PoT)

- Particularly designed to capture motion information in first-person videos
- Abstract a set of raw feature descriptors from each video into a single vector representing the video
- Result are served as an input vector for classifiers

Pipeline of PoT



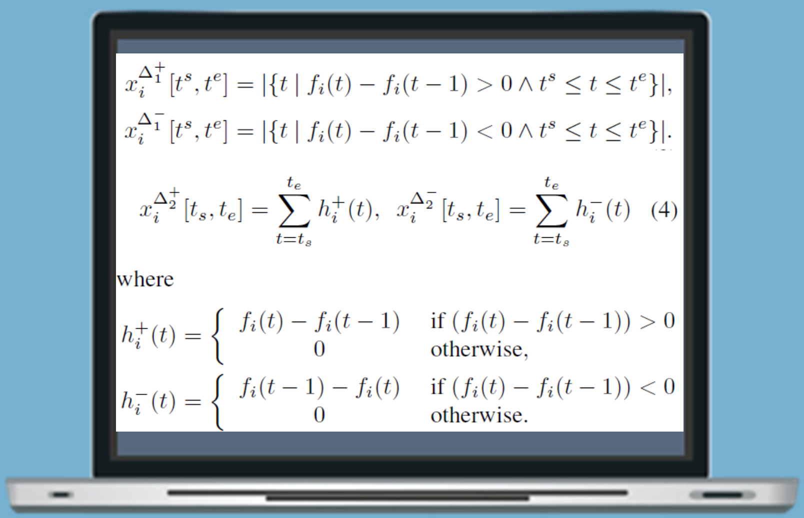
Three Important abilities of PoT

1. Allows the representation to capture both *long-term* motion and *short-term* information with multiple temporal filters.

2. Explicitly imposes *temporal structure* of the activity by decomposing the entire time interval to multiple subintervals

3. Takes advantage of multiple types of pooling operators so that the representation captures *different aspects* of the data.

Temporal pooling operators



$$x_i^{\Delta_1^+}[t^s,t^e] = |\{t \mid f_i(t) - f_i(t-1) > 0 \wedge t^s \le t \le t^e\}|,$$
$$x_i^{\Delta_1^-}[t^s,t^e] = |\{t \mid f_i(t) - f_i(t-1) < 0 \wedge t^s \le t \le t^e\}|.$$

$$x_i^{\Delta_2^+}[t_s,t_e] = \sum_{t=t_s}^{t_e} h_i^+(t), \quad x_i^{\Delta_2^-}[t_s,t_e] = \sum_{t=t_s}^{t_e} h_i^-(t) \quad (4)$$

where

$$h_i^+(t) = \begin{cases} f_i(t) - f_i(t-1) & \text{if } (f_i(t) - f_i(t-1)) > 0 \\ 0 & \text{otherwise,} \end{cases}$$
$$h_i^-(t) = \begin{cases} f_i(t-1) - f_i(t) & \text{if } (f_i(t) - f_i(t-1)) < 0 \\ 0 & \text{otherwise.} \end{cases}$$

Max Pooling

Sum Pooling

Histogram of time series gradients

number of positive (and negative) gradients within the temporal filter.

Histogram of time series gradients

sum of the amount of positive (or negative) gradients.

Representation Implementation

HOF	MBH	Overfeat CNN	Café CNN
Optical flow based motion descriptor	Optical flow based motion descriptor	Descriptors from CNNs pre-trained on ImageNet	Descriptors from CNNs pre-trained on ImageNet
200-D	400-D	4096-D	4096-D
L1 normalization applied	L1 normalization applied	L1 normalization applied	L1 normalization applied

Experimental Settings

DogCentric Dataset

wearable cameras mounted on dogs' back. ego-actions of the dog & interactions between the dog and humans

UEC Park Dataset

a human wearing a camera. ego-actions of the person (wearing a camera) involved in physical activities

Classifier

nonlinear
SVM with a X^2 kernel

Evaluation setting

Repeated random training/testing splits 100 times was performed, and averaged the performance. randomly selected half of videos per class as training videos

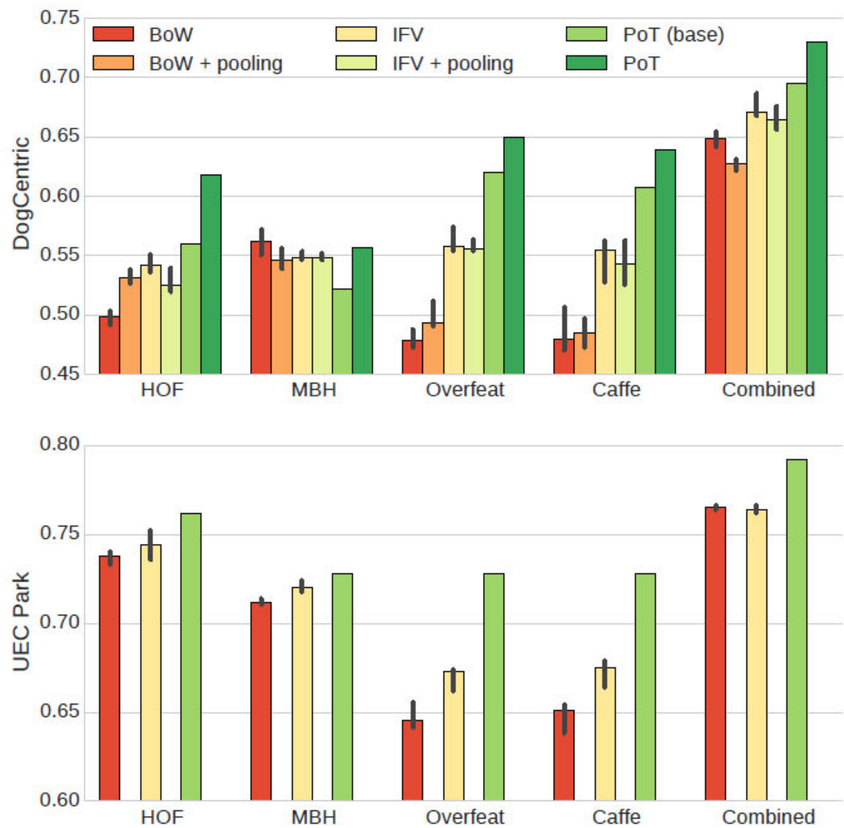


Figure 3. Classification accuracies of feature representations with each descriptor (and their combination). Representations that utilize randomness are drawn with 95% confidence intervals. See text for details.

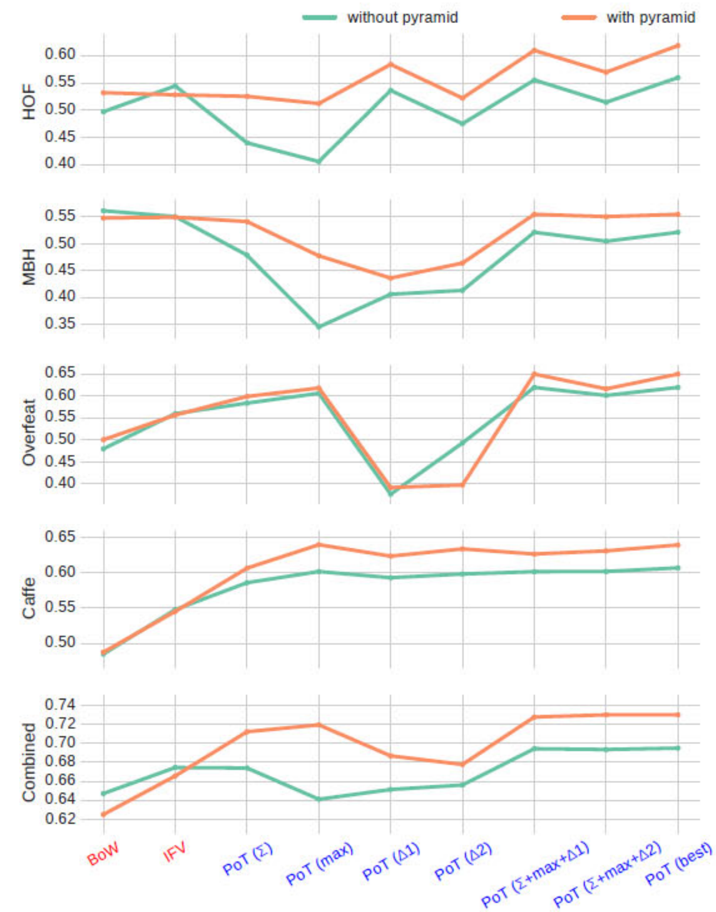


Figure 4. Feature performance using BoW and IFV compared with various PoT pooling operators with and without a temporal pyramid on the DogCentric dataset. Y-axis is classification accuracy, and X-axis shows different representations. PoT generally benefits

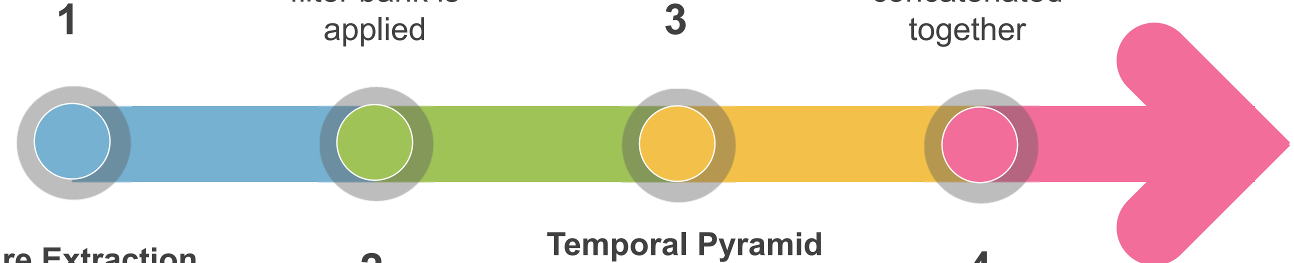
End of Part 1

Order-aware convolutional pooling

Frame-level representation

- Video frame represented by concatenating both the appearance features and the motion features.
- **Appearance Features:**
  4096-D activations of the second fully layer of AlexNet pre-trained on ImageNet
- **Motion Features:**
  Improved dense trajectory (IDT),
  trajectories falling into a local neighborhood (10 frames) considered & encoded using Fisher vector coding

Order-aware Convolutional Pooling



1

Feature Extraction

Extract frame appearance/motion descriptors.
Treat as n time series (1D signals)

2

1D Convolution

For each 1D signal, a unique filter bank is applied

3

Temporal Pyramid Pooling

For the i th level, filter responses are partitioned into m_i segments. within each segment max pooling is performed. Results of each segment will be concatenated

4

Concatenation

the dimension-level representations are concatenated together

Order-aware convolutional pooling action recognition

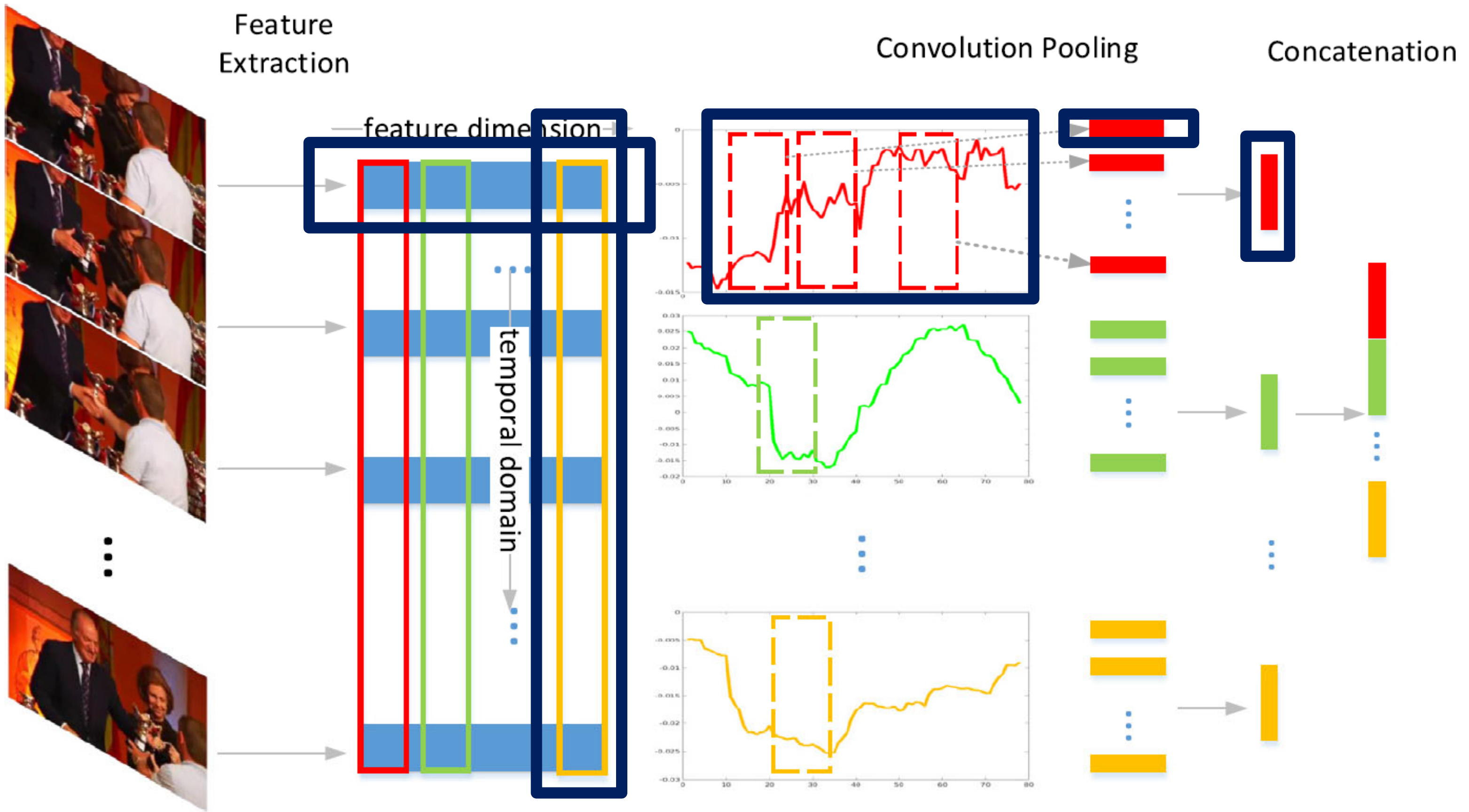


Fig. 2. Illustration of order-aware pooling.

1D Convolution for each Dimension

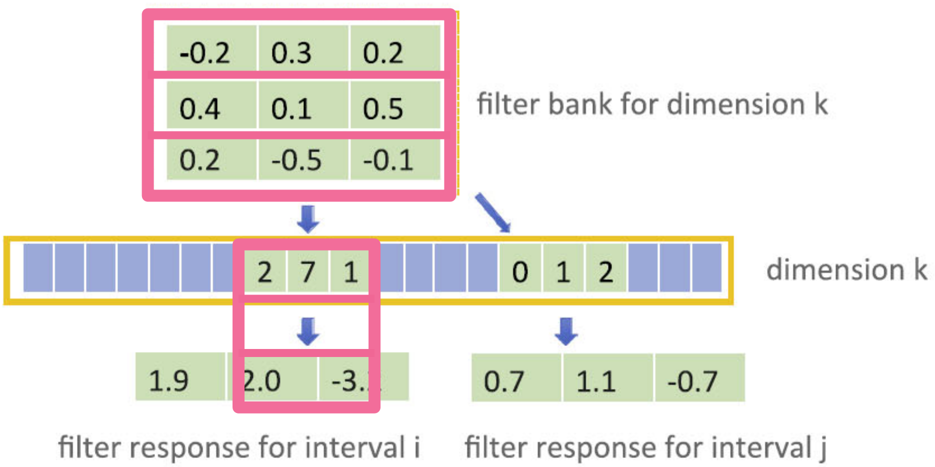


Fig. 3. Illustration of 1D convolution. In this figure, the number of filters are 3 and the interval size is 3 as well.

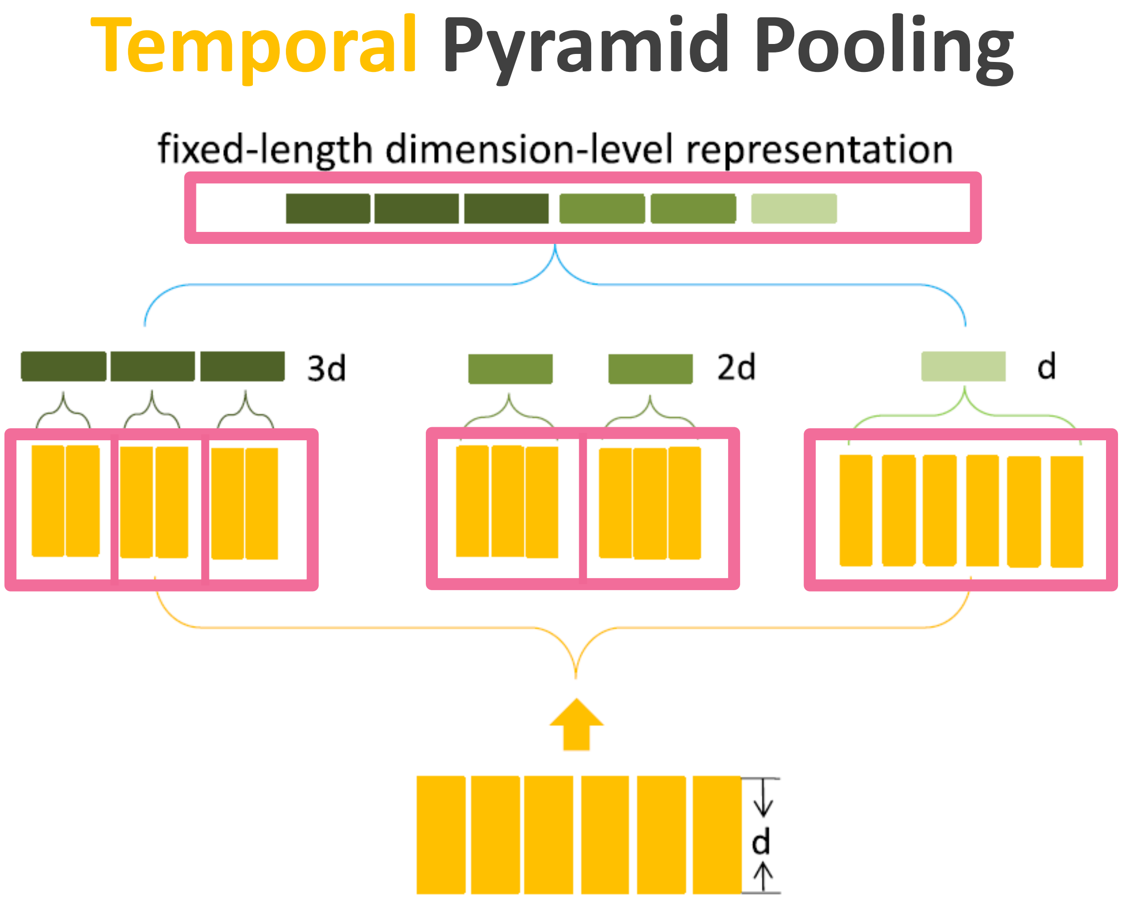


Fig. 4. Illustration of temporal pyramid pooling.

$$L(\mathbf{W}, \mathbf{b}) = -\sum_{i=1}^{N} \log(\mathbf{Y}(c_i)),$$

where c_i denotes the class label of the ith video and N is the total number of training videos. Recall that $\mathbf{Y}$ is a c-dimensional vector and c equals to the number of classes.

Classification and model parameter learning

Classifier

classification layer on top of the outputs of the pooling layer. model parameters will be learned in a supervised fashion

Optimizer

Model parameters will be updated Using Stochastic Gradient Descent (SGD)

Final Classification

Final classification scores are calculated as sum of classification scores from appearance features and motion features.

Experimental setup

- **HMDB51 dataset**
  Collected from various sources such as YouTube
  6766 video clips, 51 classes
- **UCF101 dataset**
  Realistic action videos collected from YouTube
  13,320 videos, 101 classes.
- **Hollywood2 dataset**
  1707 videos with 823 training and 884 testing videos
  69 movies, 12 classes

Comparisons

Table 1
Comparison of the proposed pooling method to the baselines on HMDB51 using appearance information or motion information.

Appearance	AP	37.5%
	MP	36.5%
	PoT (no TP) [46]	36.5%
	TP	39.2%
	Ours (MP)	**40.8%**
	Ours (TP)	**41.6%**
Motion	AP	50.9%
	MP	50.6%
	TP	54.7%
	Ours (MP)	52.8%
	Ours (TP)	**55.0%**

Table 2
Comparison of the proposed pooling method to the baselines on UCF101 using appearance information or motion information.

Appearance	AP	66.3%
	MP	67.4%
	PoT (no TP) [46]	67.5%
	TP	68.5%
	Ours (MP)	**69.3%**
	Ours (TP)	**70.4%**
Motion	AP	80.0%
	MP	80.2%
	TP	81.6%
	Ours (MP)	81.0%
	Ours (TP)	**82.1%**

Conclusion

- Convolution operation constitutes the most important part of the proposed pooling method.
- Motion features can lead to better classification performance comparing to appearance features.
- On appearance features, the proposed pooling method can con- sistently outperform the baselines

Thank you

I would like to express my sincere gratitude to my supervisor Dr. A. Mansouri