Multimedia Event Detection Using Segment-Based Representation

Sang Phan

The Graduate University for Advanced Studies (SOKENDAI)
plsang@nii.ac.jp

May 20th, 2014

Table of Contents

Table of Contents

Multimedia Event Detection

Motivation

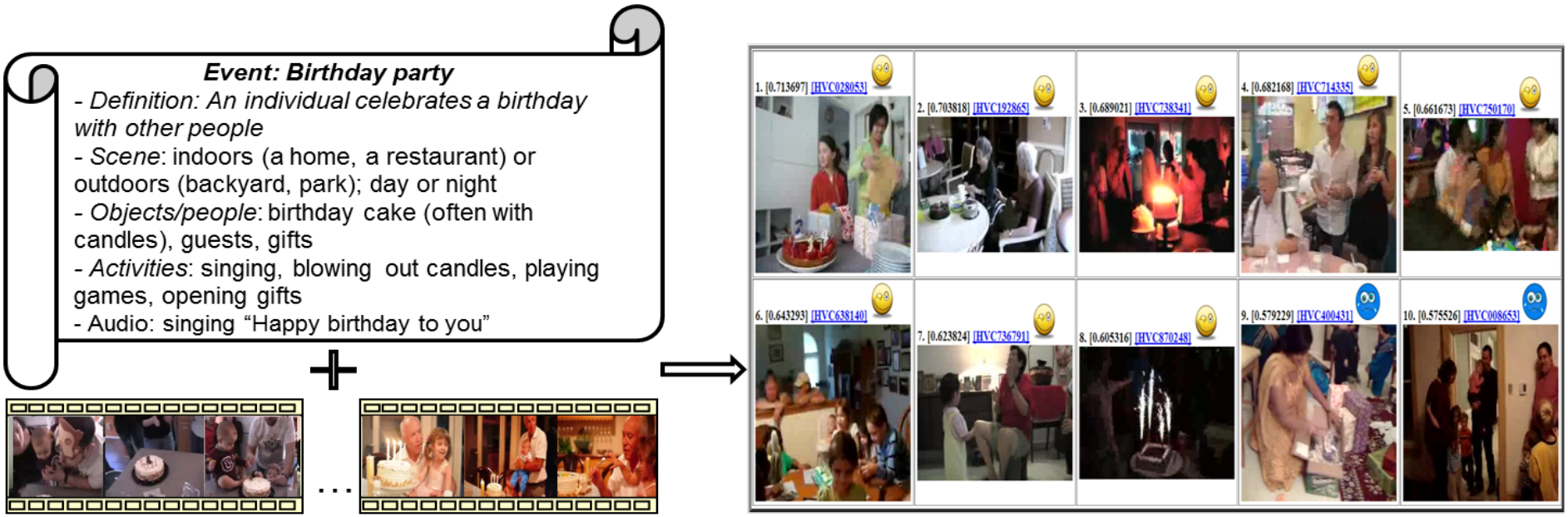
- Massive number of videos are produced every day.
  - YouTube: 72 hours uploaded per minute, with 3 billion viewers a day.
- Video need to be indexed, searched based on its content.
- Many applications:
  - User demands: tutorial videos such as "**how to make a cake**", "**how to repair an appliance**".
  - Security purposes: filter out irrelevant content such as "**how to make a bomb**".

Multimedia Event Detection

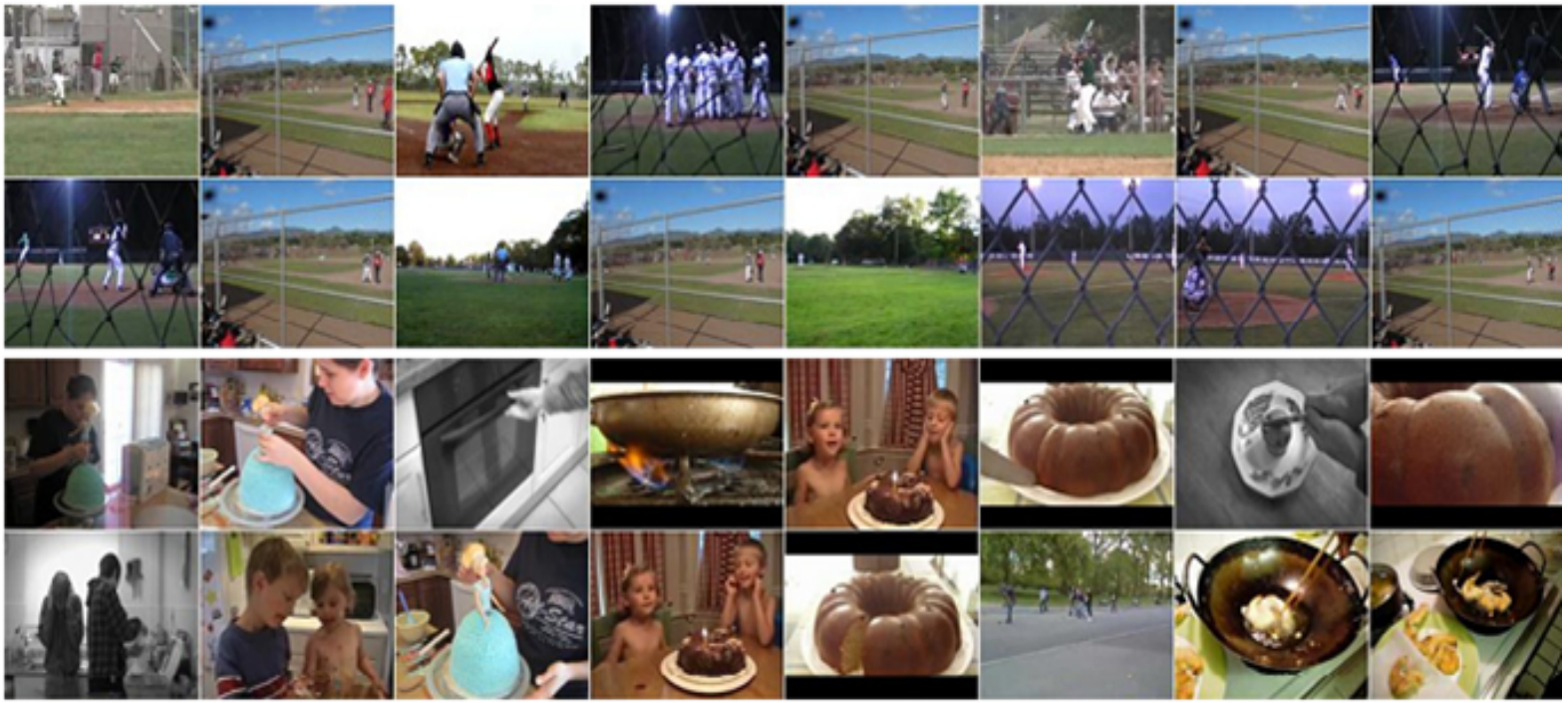
Task defined by TRECVID since 2010

Definition

- Given: An event kit which consists of an event name, definition, explication + video example.
- Wanted: A system that can search for this event through the large set of videos with reasonable accuracy and speed.



Challenges of Multimedia Event Detection



- Large content variation: Large number of events and large number of background videos.
- Uncontrolled capturing conditions: different time, location, clutter in the environment, camera motion.

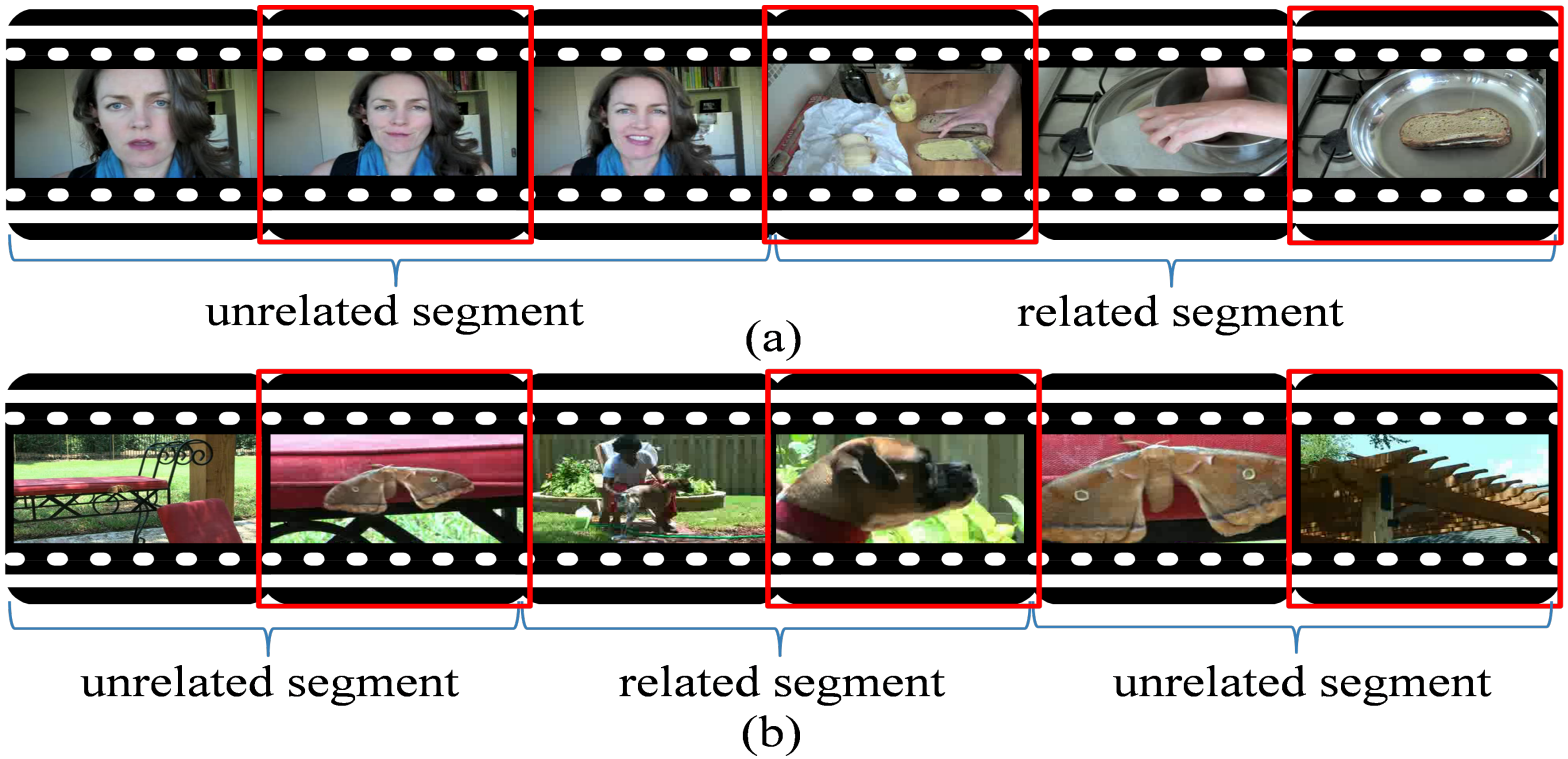
Challenges of Multimedia Event Detection

- Evaluation datasets:

Dataset	MED 2010	MED 2011	MED 2012
Number of test events	**3** (Assembling a shelter, Batting a run, Making a cake)	**10** (Birthday party, Changing a vehicle tire, Flashmob gathering, etc)	**20** (Cleaning an appliance, Dog show, Marriage proposal, etc)
Number of videos	**3,468** (1,744 dev videos and 1,724 test videos)	**45,000** (13,200 dev videos and 31,800 test videos)	**156,000** videos (58,000 dev videos and 98,000 test videos)
Number of background videos	**1,500** for dev and **1,500** for test	**10,000** for dev and **28,000** for test	**10,000** for dev and **95,000** for test
Hours of video	**110**	**1,400**	**4,850**

Challenges of Multimedia Event Detection

- **Specific challenge**: Data often contain irrelevant information

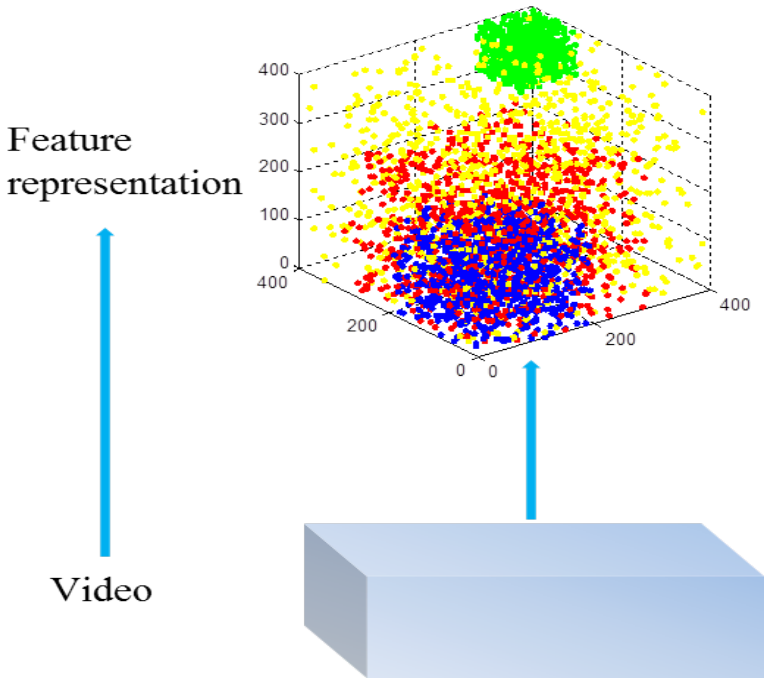


(a) Example video for "making a sandwich" event: the related segment appears after a self-cam segment (unrelated); (b) example video for "grooming an animal" event: related segment is sandwiched between two unrelated segments. This kind of video is popular in realistic dataset.

Table of Contents

Video-based Approach

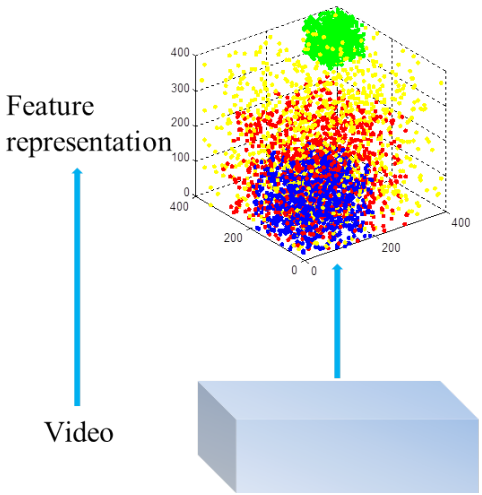
- Features are computed over the whole video
- One representation for each video



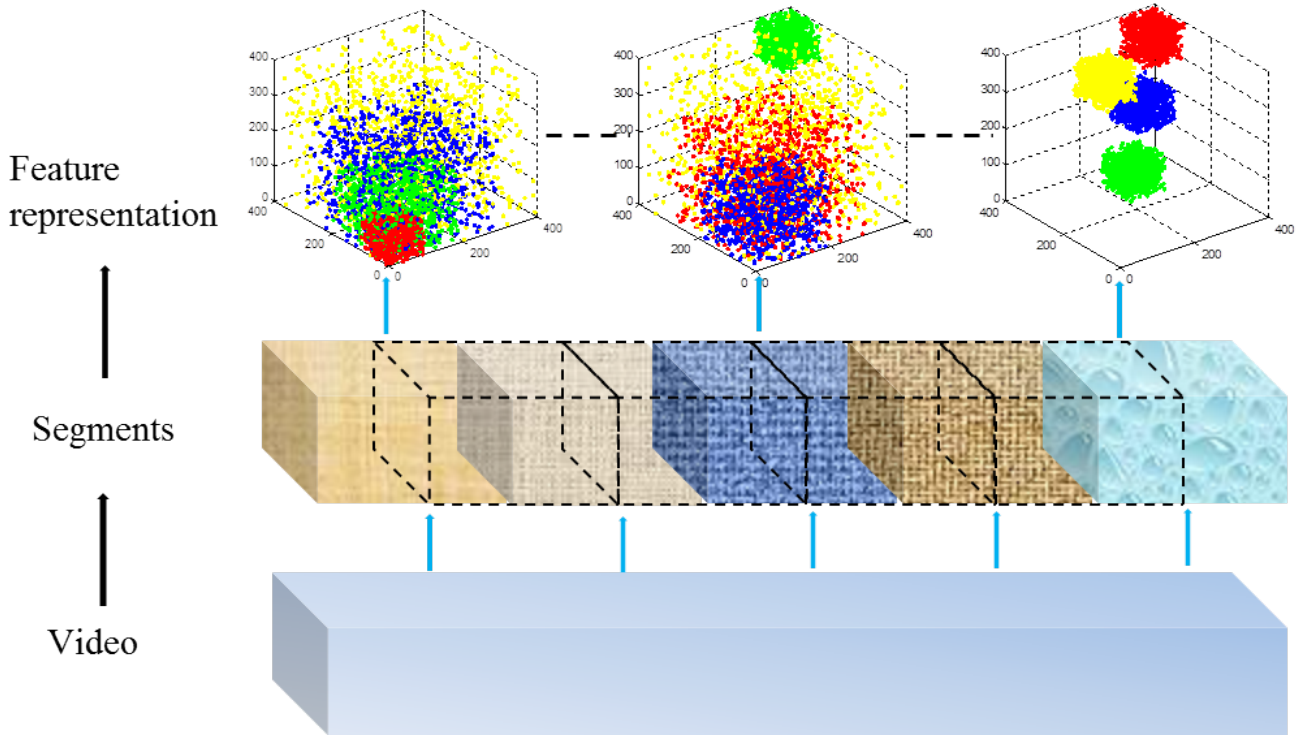
- Used by best MED'10 system (Columbia University)
- Used by best MED'11 system (BBN VISER)

Specific problem: The clues to determine an event can reside within a small segment.

Our Segment-based Approach



(b) The video-based approach



(b) **Our proposed segment-based approach:** the basic idea is to examine shorter segments instead of using the entire video

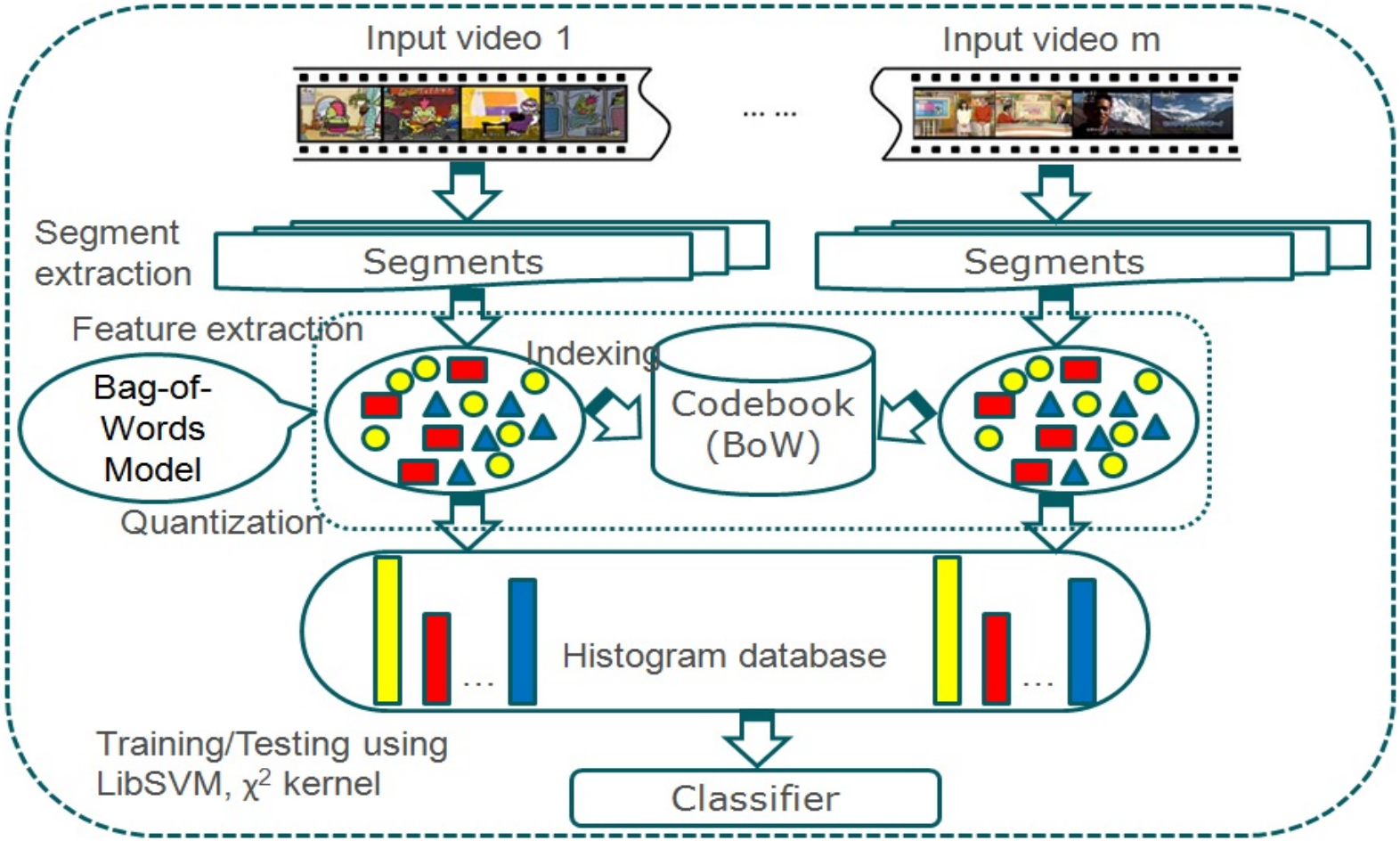
Our Segment-based Approach

How to select the segment length?

- Non-overlapping
  - Uniform sampling
  - Segment length: 30, 60, 90, 120, 200, 400 seconds
  - Compare with the video-based approach (using the whole video)
- Overlapping sampling
  - Uniform sampling, 50% overlapping
  - Segment length: 30, 60, 90, 120, 200, 400 seconds
  - Compare with the video-based approach (using the whole video)
- Segment sampling based on shot boundary detection
  - Take into account the boundary information of each segment
  - Employ the technique proposed by [Guimaraes et al. - 2003]

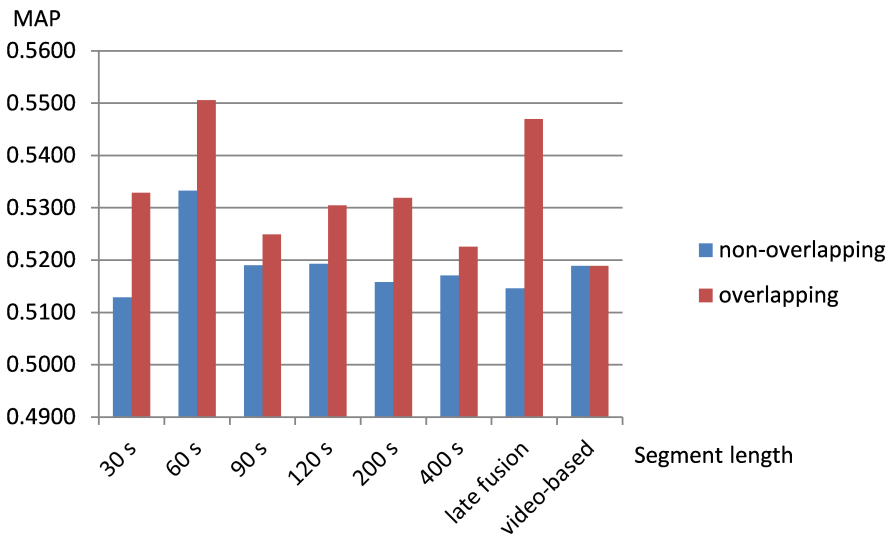
Guimaraes, S.J.F., Couprie, M., Araujo, A.d.A., Leite, N.J: Video segmentation based on 2d image analysis. Pattern Recognition Letters, 2003, 24(7), 947957.

Evaluation Framework

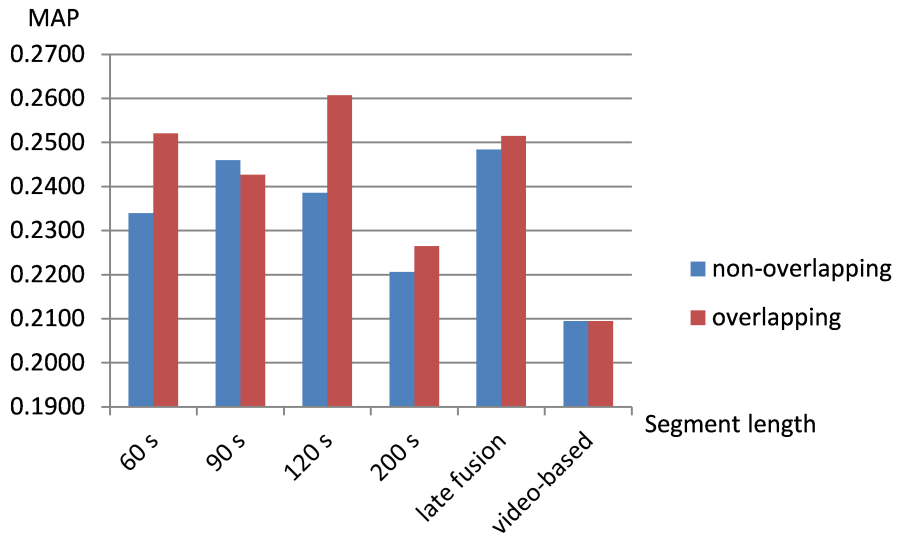


Evaluation framework for our baseline MED system

Result: Non-Overlapping vs. Overlapping Sampling



(b) On the MED 2010 dataset



(b) On the MED 2011 dataset

In most cases, the overlapping sampling performs the best.

Result: Comparison

Table: Comparison of different segment-based approaches with the video-based approach on the MED 2010 dataset.

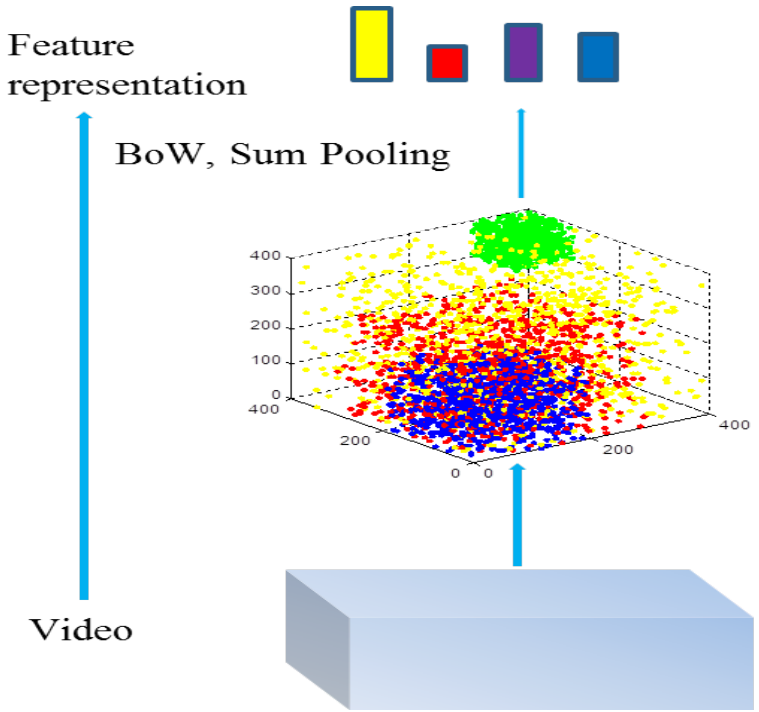
Event/MAP	Best non-overlapping	Best over-lapping	SBD seg-ments	Video-based
Assembling shelter	0.4511	0.4781	0.4284	**0.4911**
Batting in a run	0.7852	**0.7918**	0.7866	0.7902
Making a cake	0.3636	**0.3819**	0.1918	0.2755
All	0.5333	**0.5506**	0.4689	0.5189

Segment-based approach outperforms the traditional video-based approach.

Table of Contents

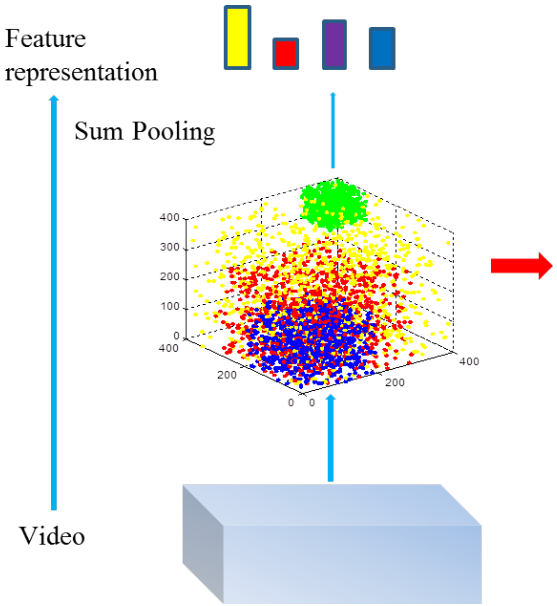
Video-based Approach

Bag-of-visual-words model: Video level features are aggregated over the entire videos

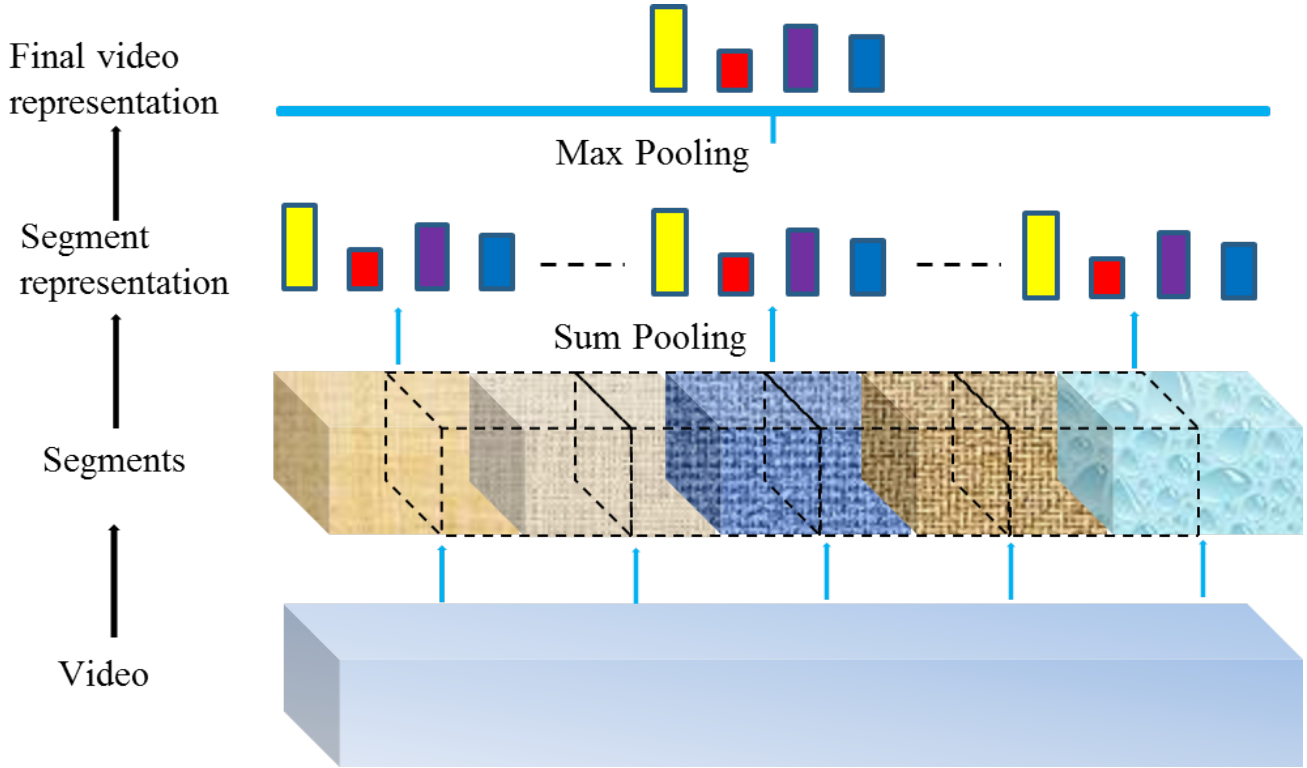


- The proposed segment-based approach: increasing the number of segment representation is not scalable.
- **Specific problem:** How to generate one representation per each video from its segment-level representations?

Sum-max Video Pooling



(b) The video-based approach



(b) **Our proposed Sum-max Video Pooling:** the basic idea is to retrieve the most relevant features at the high layer, therefore it can discard irrelevant features in the final video representation

Sum-max Video Pooling

- N local descriptors $x_n \in R^D$, where n = 1,...,N and D is the feature dimension
- K visual words $m_k \in R^D$, where k = 1,...,K
- $M = \{m_k\}$ is the set of visual words
- Coding step: $\phi_n = [\Phi_{1n}, ..., \Phi_{Kn}]$
- S is number of segments
- N_s is the number of local descriptors in segment s
- The sum-max and max-sum video pooling at each visual word can be defined as follows:

$$\psi_{k_{\text{sum-max}}} = Max_{s \in S}(\sum_{n \in N_s} \Phi_{kn}) \quad (1)$$

$$\psi_{k_{\text{max-sum}}} = \sum_{s \in S}(Max_{n \in N_s} \Phi_{kn}) \quad (2)$$

Experimental Results

Table: Performance comparison of different video pooling strategies on the MED 2010 dataset.

Event/MAP	Max pooling (Video-based)	Sum pooling (Video-based)	Max-sum pooling (at 60 s)	Sum-max pooling (at 60 s)
E001	0.4365	0.4468	0.4646	**0.5072**
E002	0.6434	**0.7988**	0.7103	0.7900
E003	**0.3144**	0.3053	0.2806	0.3100
All	0.4648	0.5170	0.4852	**0.5357**

- Pooling over segments is more effective.
- **Sum-max video pooling outperforms the traditional video-based sum pooling.**

Where am I now?

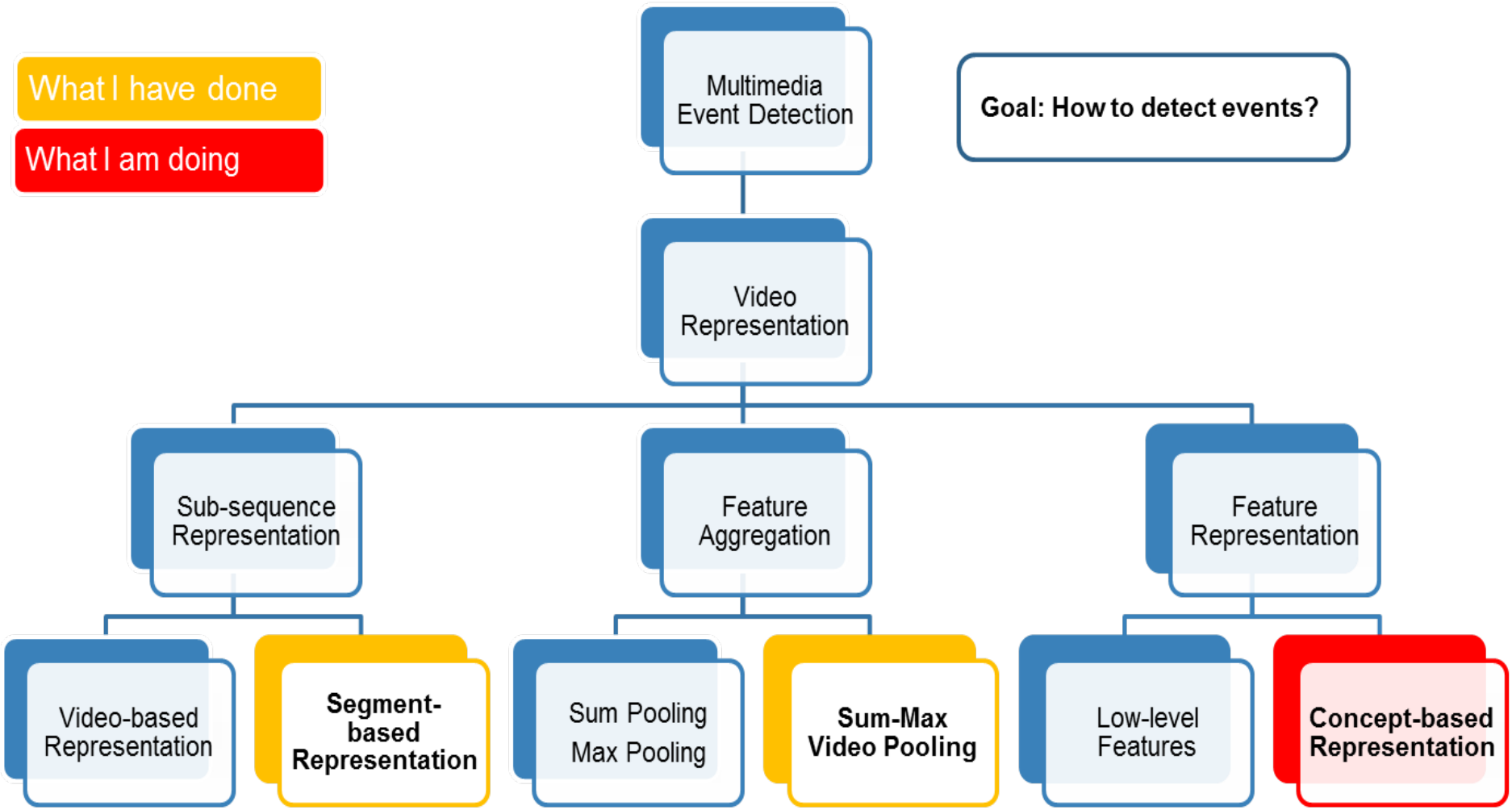
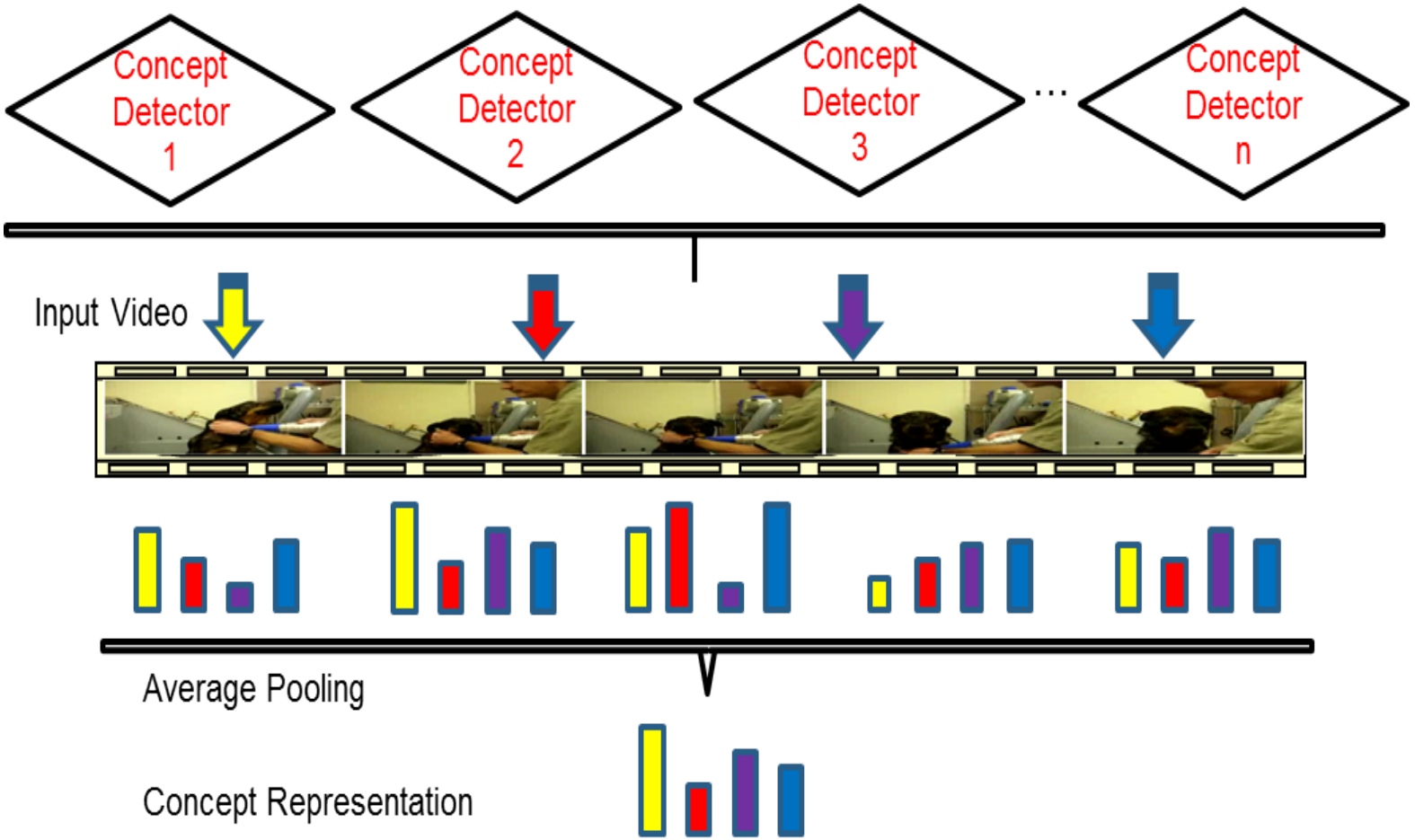


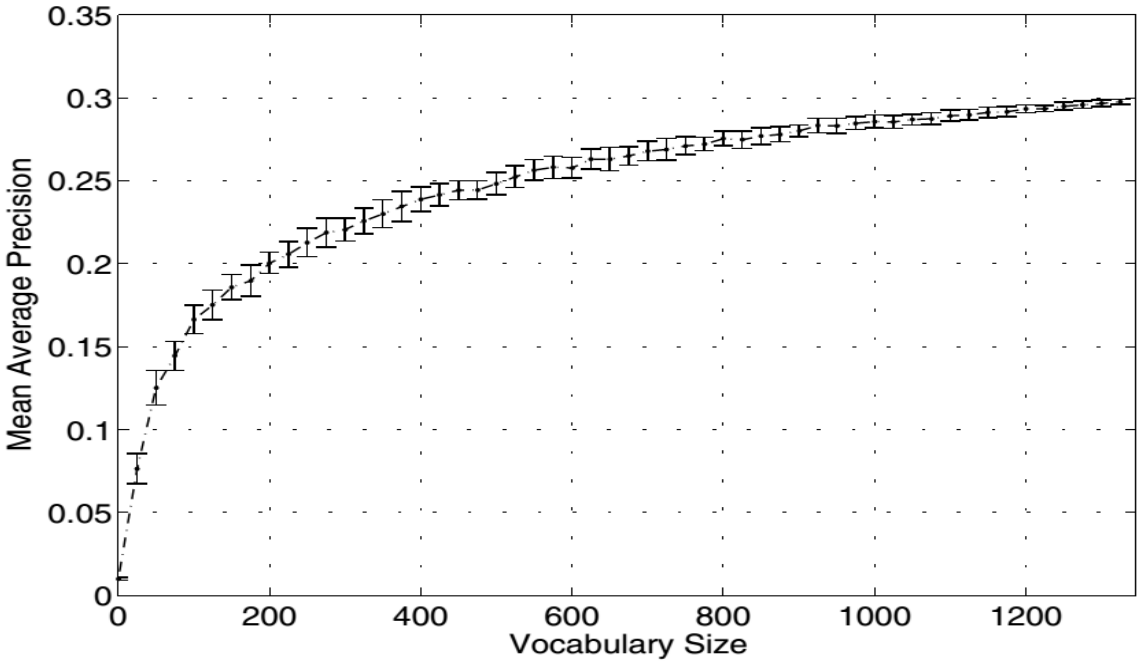
Table of Contents

Concept-based Representation

A complex event can involve multiple objects $\rightarrow$ **A video can be represented by concept vectors.**



Concept-based Representation



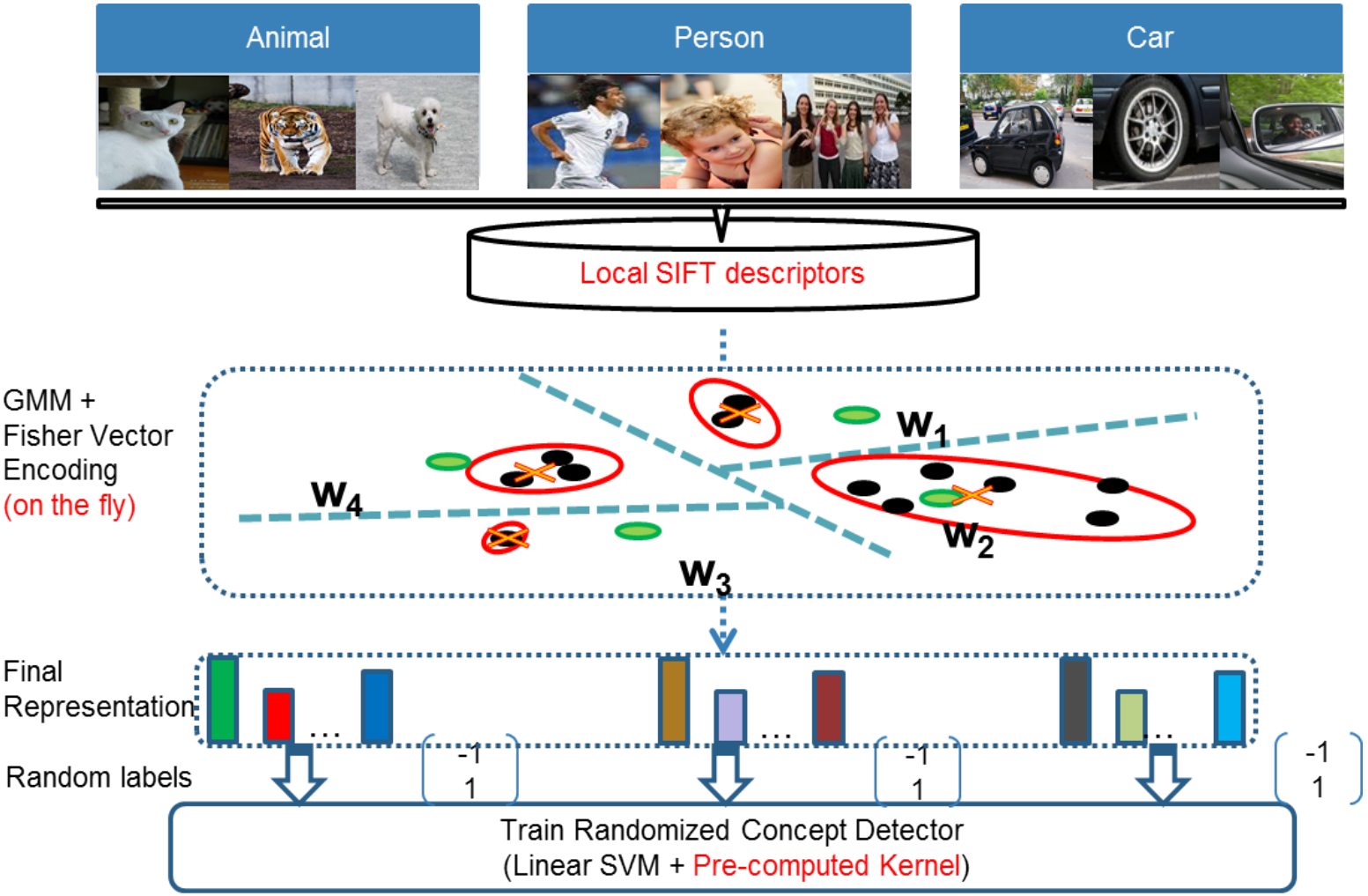
- Increasing the concept vocabularies can improve the event recognition performance. [Habibian - ICMR2013]
- Limitation: Annotation requires a lot of human effort!

AmirHossein Habibian, Koen E. A. van de Sande, Cees G. M. Snoek: Recommendations for video event recognition using concept vocabularies. ICMR 2013: 89-96

How About an Effortless Approach?

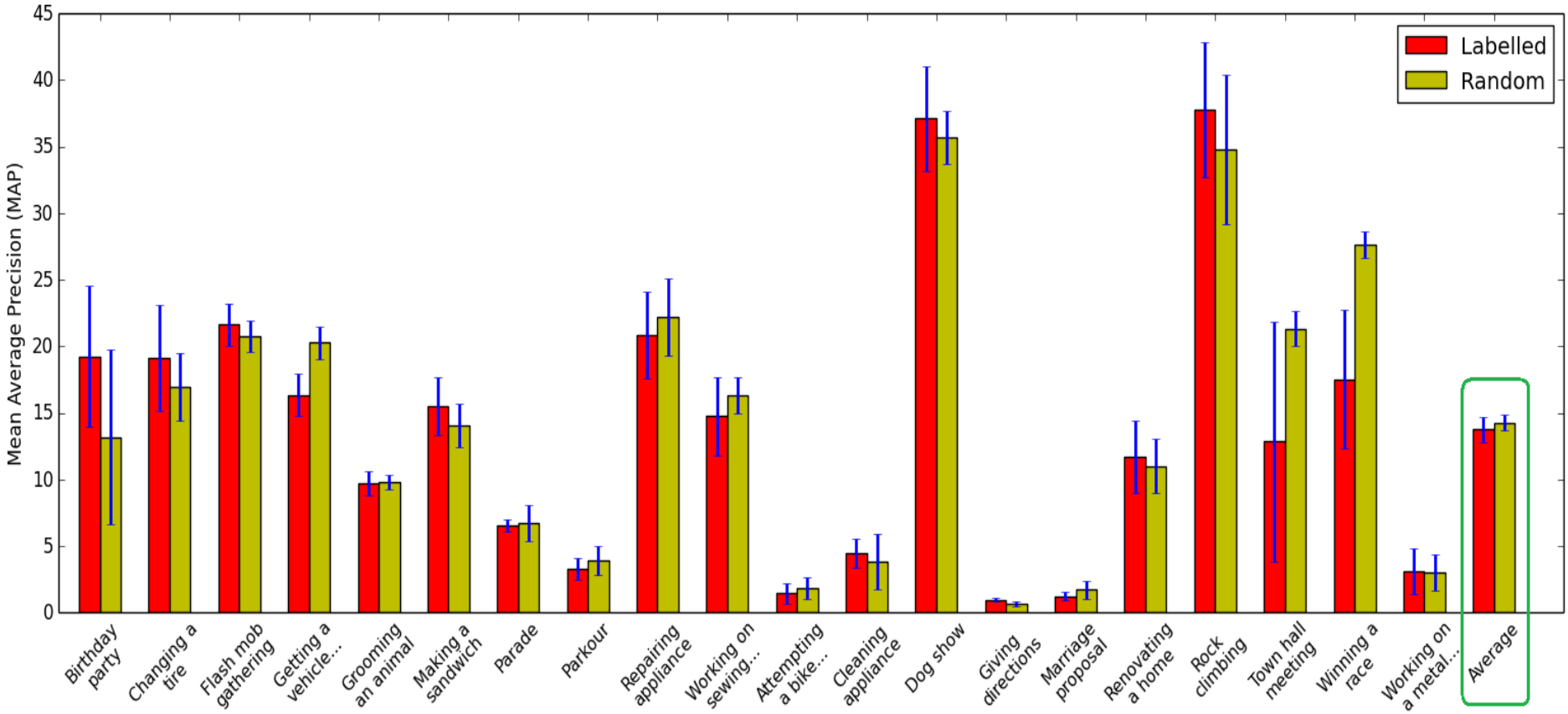
- We introduce a new approach in order to get labels for training data efficiently.
- **Random Concepts**: Given a set of images used to train a concept classifier, positive and negative labels are generated randomly.
  - In most of cases, the presence of a concept is null
  - The performance of concept detection is usually poor
  - Different domain problem

Randomized Concept Detector Framework



Framework for building randomized concept detectors from ImageNet dataset

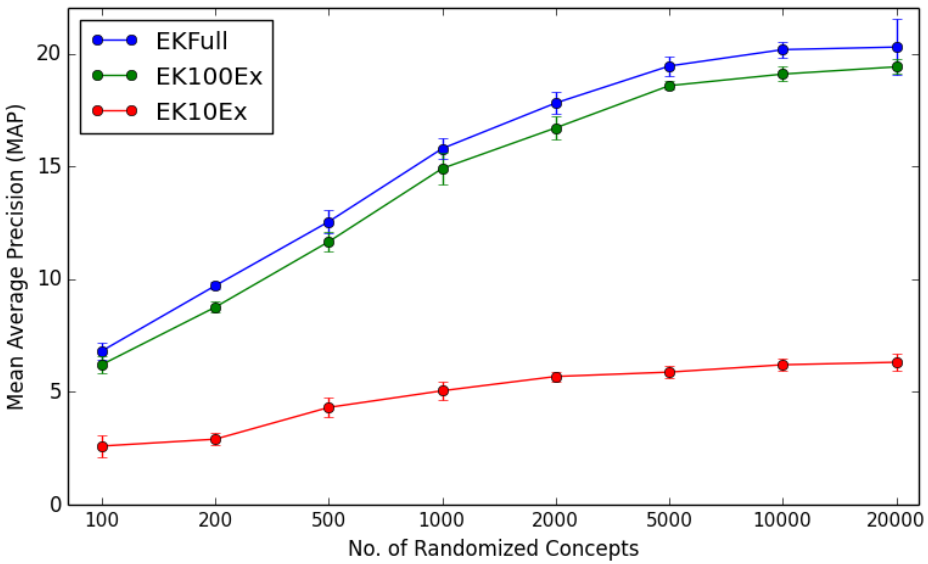
Labelled vs. Random Concept Detector?



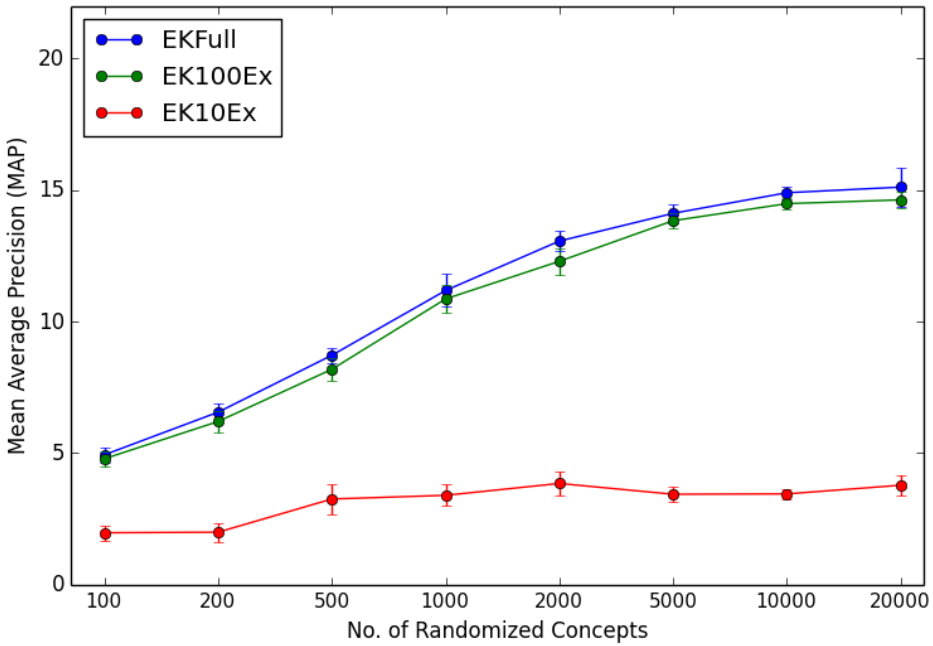
Performance of 1000 labelled concept detectors vs. Performance of 1000 random concept detectors.

We can obtain comparable performance with labelled concepts!

How Many Random Concepts?



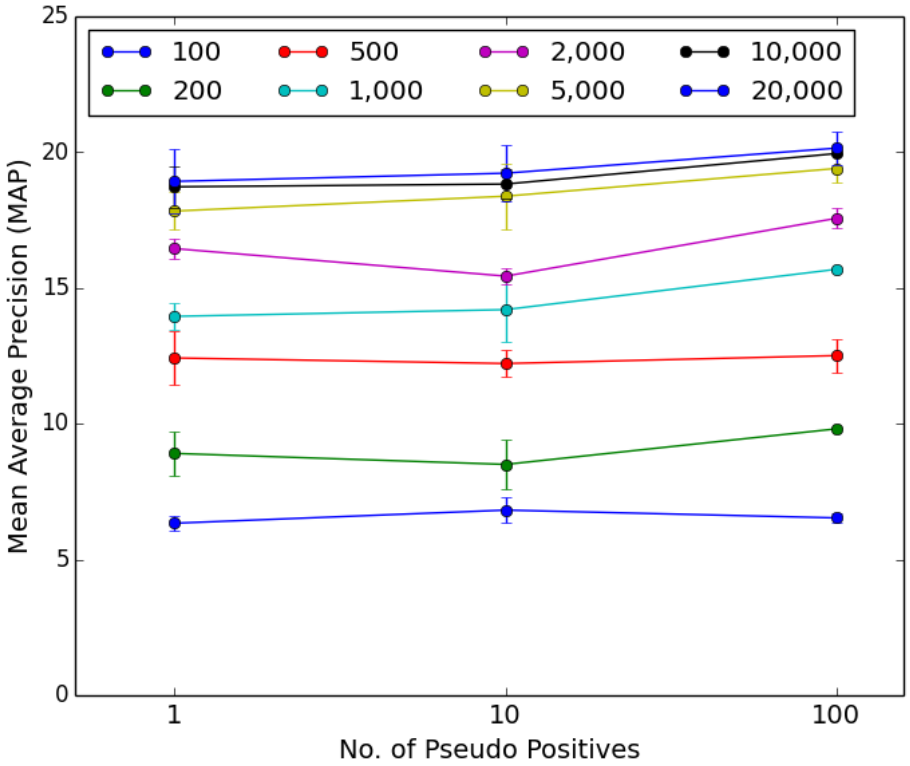
(a) On the MED 2012 KINDREDTEST dataset



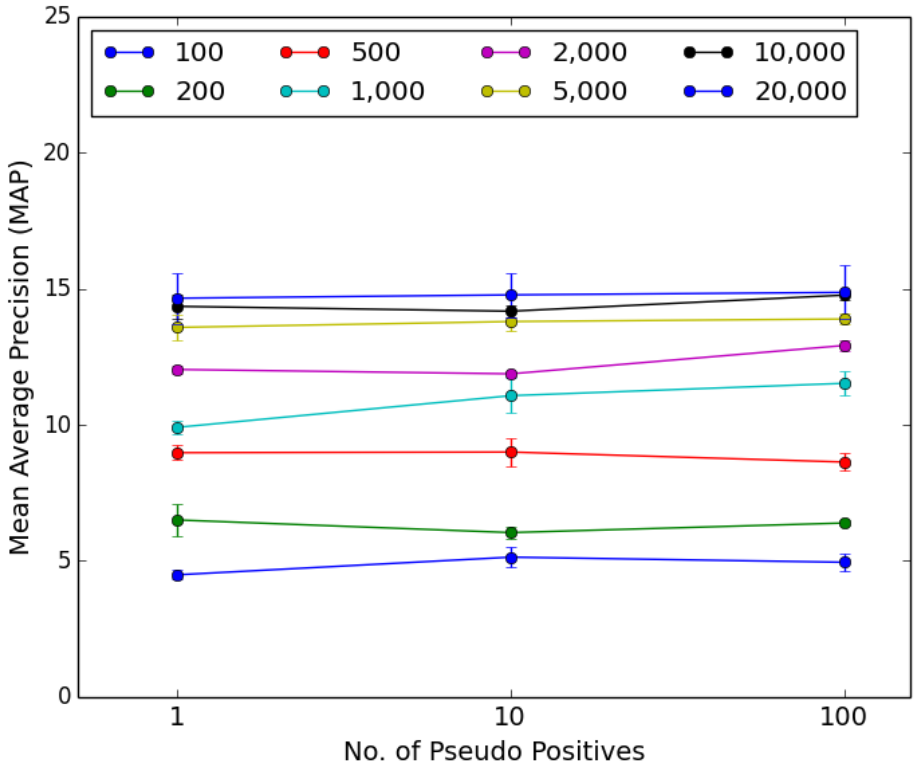
(b) On the MED 2012 MEDTEST dataset

The best performance can be obtained using around **ten thousands** of totally random labelled concept detectors!

Influence of Number of Pseudo Positives?



(a) On the MED 2012 KINDREDTEST dataset



(b) On the MED 2012 MEDTEST dataset

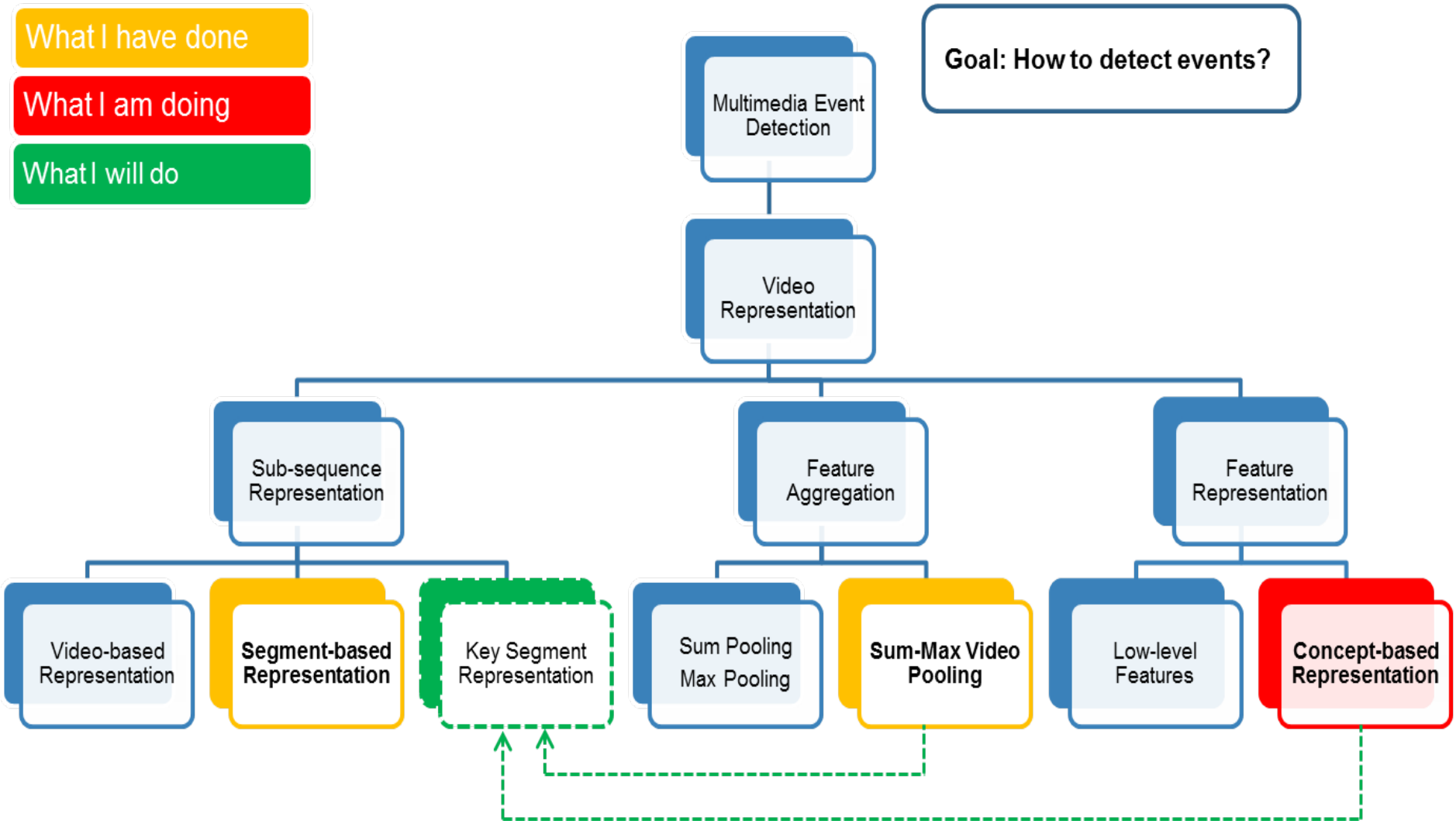
The detection performance **does not depend on the number of pseudo positives**!

Discussions

- Current results show the limitation of concept-based representation!
- Using totally random 'concept' detectors does not capture any semantic features!
- How to improve?
  - Build more concept detectors for each labelled image class?
  - How about random projection of classifiers?

Table of Contents

Next Study


What I have done
What I am doing
What I will do
Goal: How to detect events?
Multimedia Event Detection
Video Representation
Sub-sequence Representation
Feature Aggregation
Feature Representation
Video-based Representation
Segment-based Representation
Key Segment Representation
Sum Pooling
Max Pooling
Sum-Max Video Pooling
Low-level Features
Concept-based Representation

Motivation

Definition

Key segments: segments that contain positive evidence for a specific event

- Existing work addresses the problem without identification of key segments.
- Video level features are aggregated over the entire videos.

Drawback: Each part of the video contributes equally to the final representation $\rightarrow$ making it prone to noise.
For our segment-based approach: features are aggregated over the uniform sampled segments $\rightarrow$ might not contain key segments
Research Problem: How about automatically finding the key segments that contain positive evidence for a specific event?

First Approach

- We first analyze videos of specific event kit and manually identify key video segments to well represent that event type ()
  - E.g., "no dance" segment followed by "dance by group of people" segment for "flash mob gathering" event
- Train classifiers for each types of video segments (using current state of the art event detection method)
- Combine the results

Issues

- How to identify key video segments?
- How about classes of video segments?
  - Can be a set of generic classes (generic sub events)?
  - Depending of each event (sub events specific to each event)?
- How to combine the analysis results of video segment?

Summary

What I have done

What I am doing

What I will do

Goal: How to detect events?

Multimedia Event Detection

Video Representation

Sub-sequence Representation

Feature Aggregation

Feature Representation

Video-based Representation

Segment-based Representation

Key Segment Representation

Sum Pooling

Max Pooling

Sum-Max Video Pooling

Low-level Features

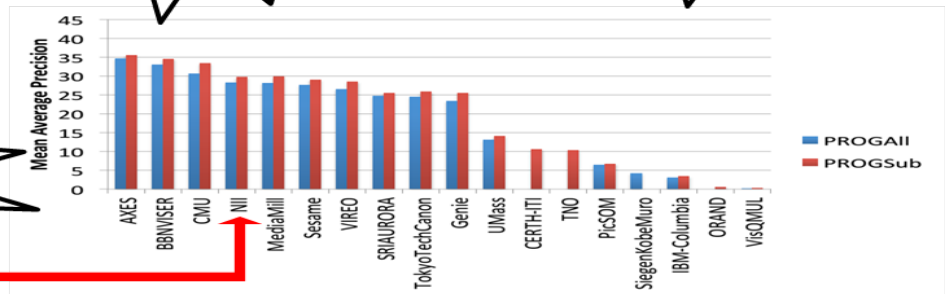
Concept-based Representation

PCM 2012
JSPS 2014
(published)

ICIP 2014
(submitted)

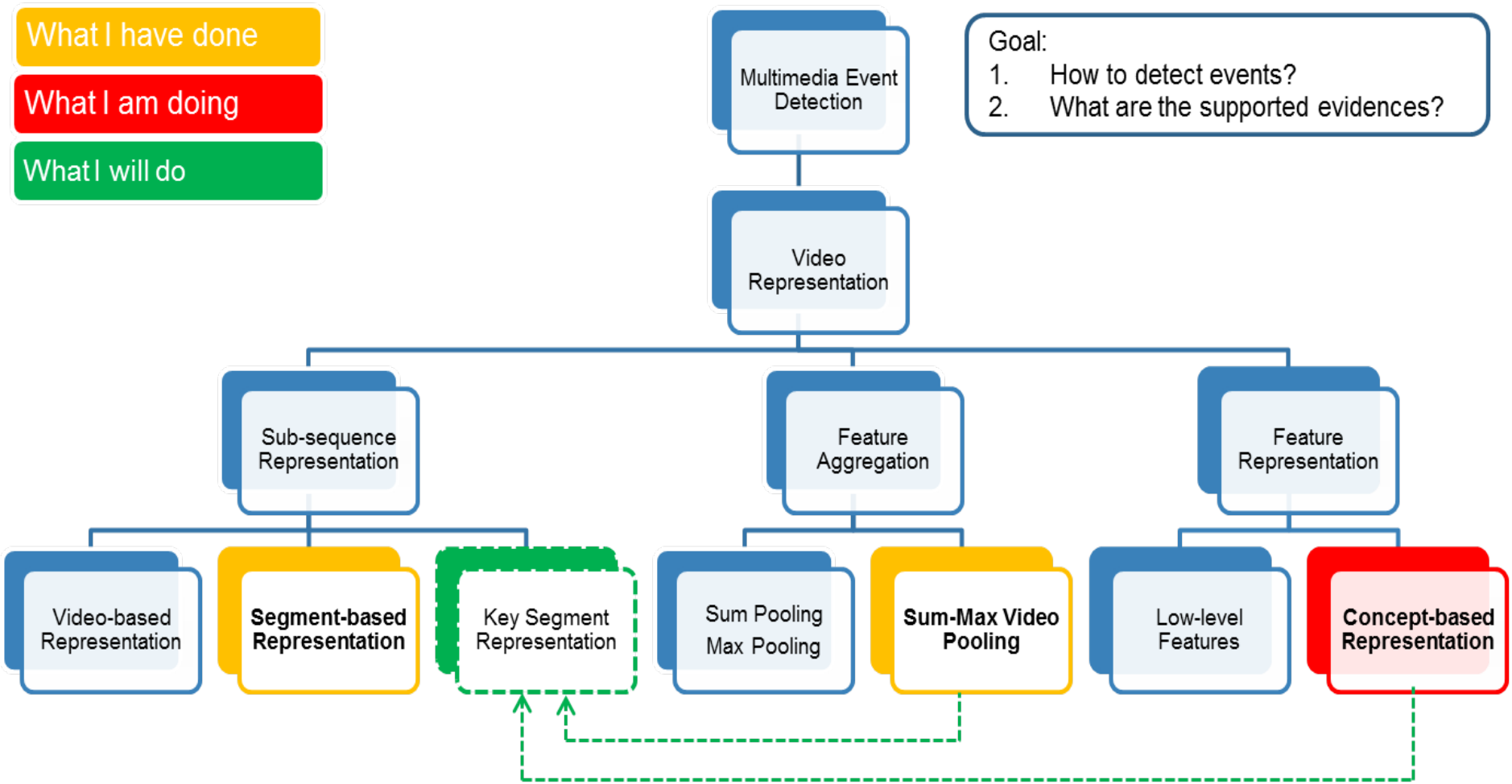
BMVC 2014
(submitted)

MED'13 Results
Ranked 4/18



Thank you for your attention!

Big Picture


What I have done
What I am doing
What I will do
Multimedia Event Detection
Goal:
1. How to detect events?
2. What are the supported evidences?
Video Representation
Sub-sequence Representation
Feature Aggregation
Feature Representation
Video-based Representation
Segment-based Representation
Key Segment Representation
Sum Pooling
Max Pooling
Sum-Max Video Pooling
Low-level Features
Concept-based Representation

Challenges of Multimedia Event Detection

- Event name: Making a cake
- Event definition: One or more people make a cake
- Event description:
  - *scene*: indoors, typically a kitchen in a home, restaurant or other setting
  - *objects / people*: ingredients, bowls, spoons, mixers, etc.
  - *activities*: selecting ingredients, combining ingredients, pouring batter into pan, putting cake into oven,
- Few video examplars

- Searching for ad hoc events
  - Limited number of video examples
  - Textual descriptions are readily understood by humans, difficult to encode for machine learning approaches

Challenges of Multimedia Event Detection

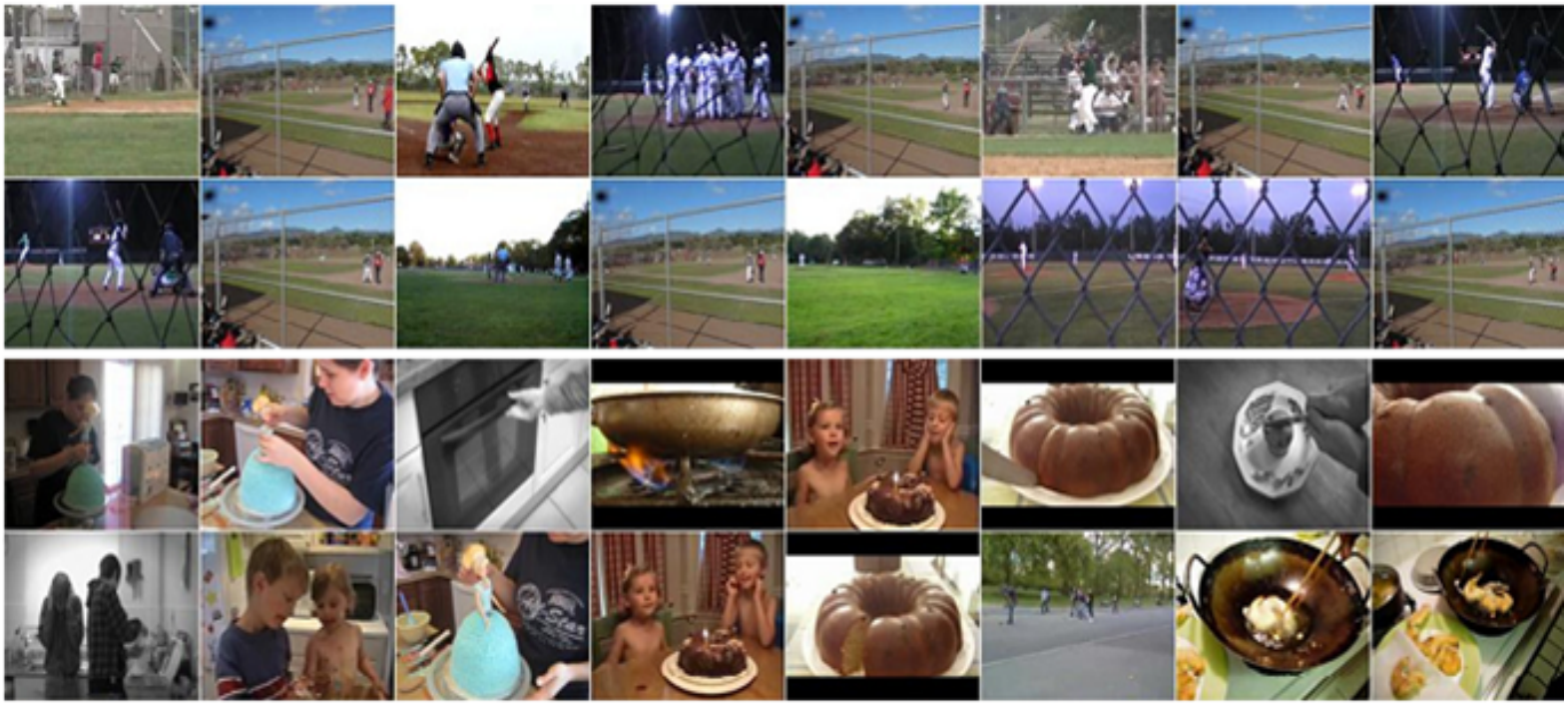
Making a cake

Batting a run in

Assembling a shelter

- Large content variation
  - Large number of events
  - Large number of background videos.

Challenges of Multimedia Event Detection

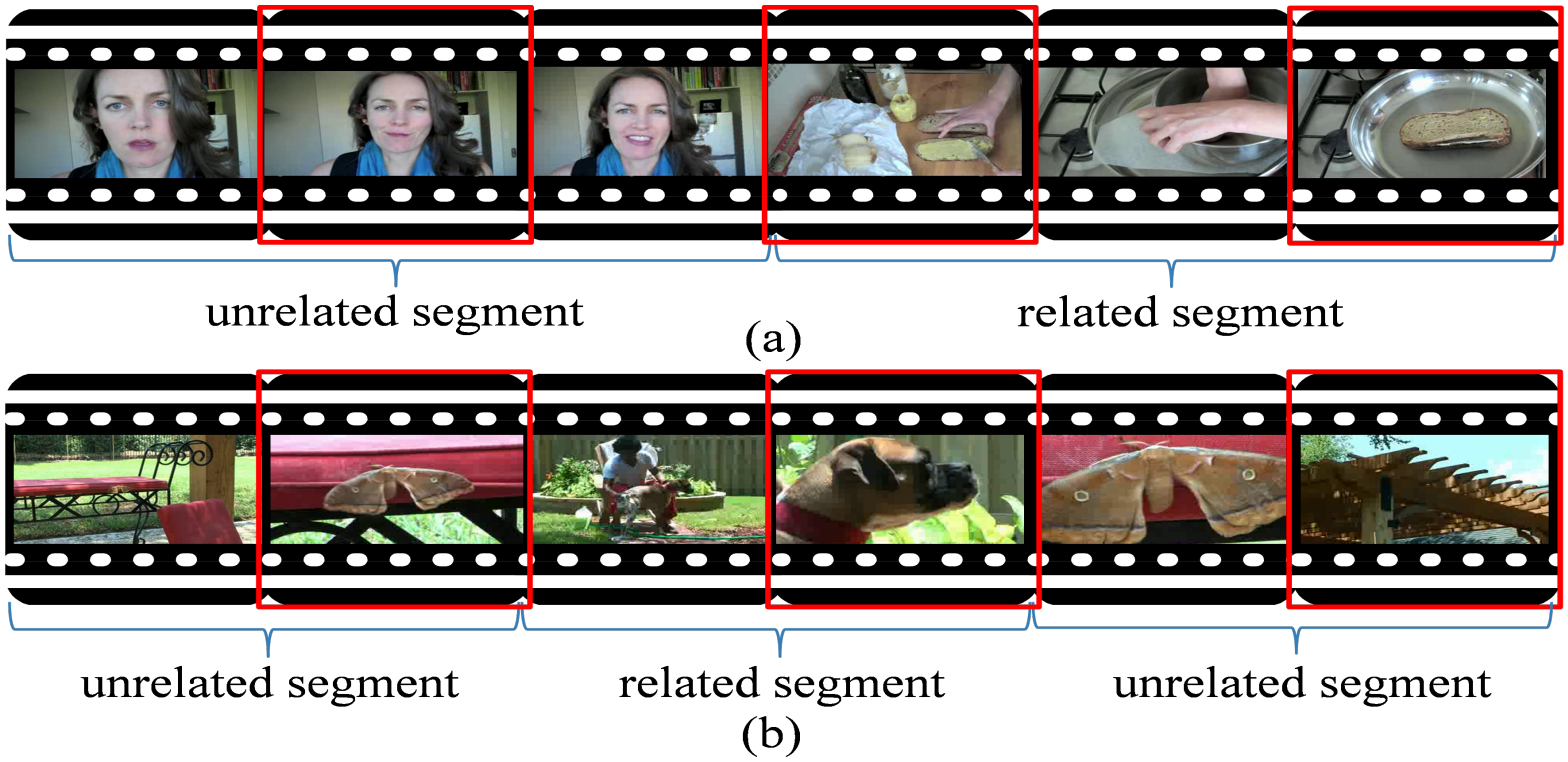


- Uncontrolled capturing conditions
  - Different time, location,
  - Clutter in the environment, camera motion, etc.

Approaches for MED

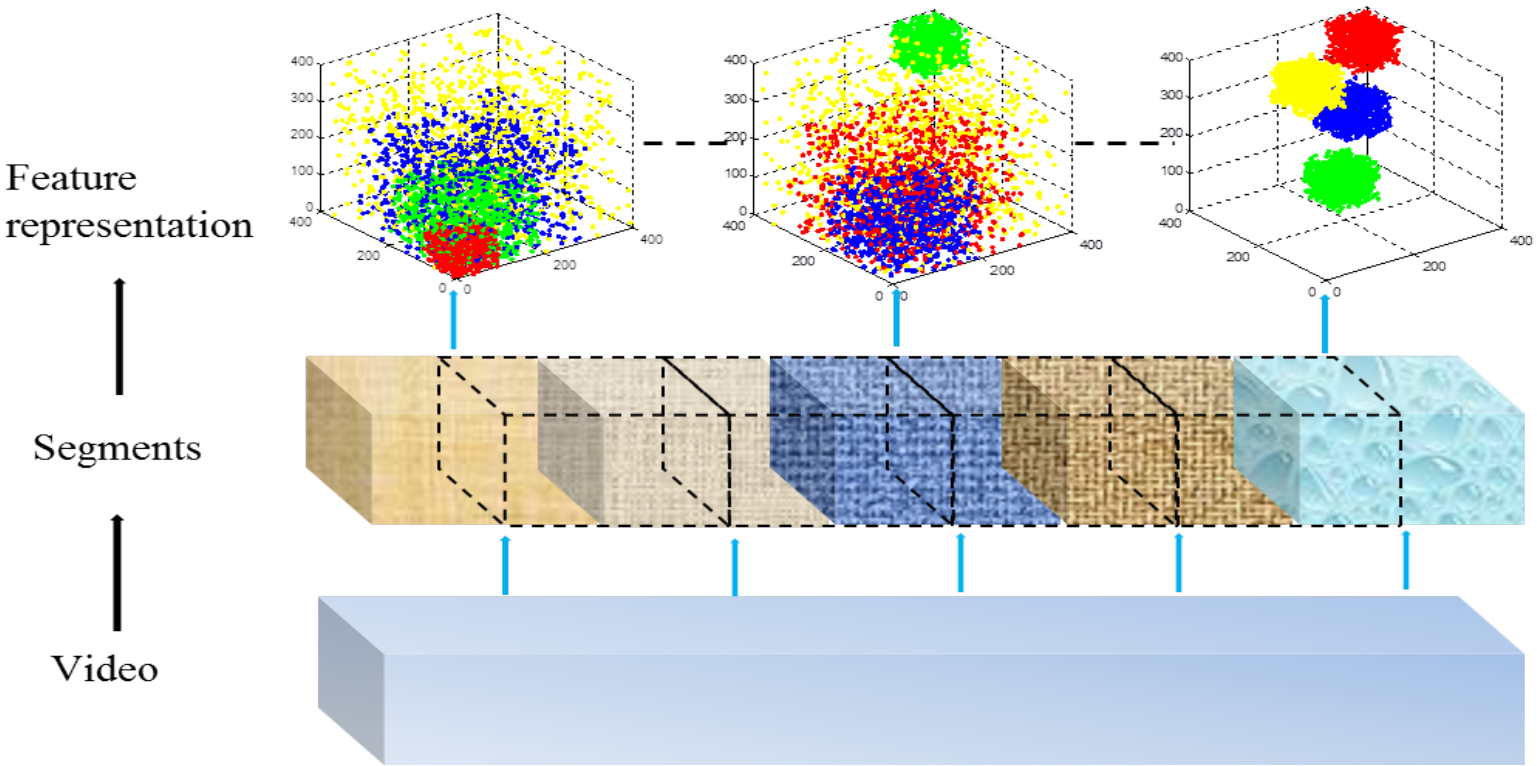
- Best MED'10 system: Columbia University, USA
  - Image features: SIFT [Lowe 2004]
  - Motion features: STIP [Laptev 2004]
  - Audio features: MFCC
- Best MED'11 system: BBN VISER, USA
  - Image features: SIFT, SURF, D-SIFT, CHOG, RGB-SIFT
  - Motion features: STIP, D-STIP
  - Audio features: MFCC, FDLP
- Common Approach: Combining multiple modalities (image, video, audio, etc.)
  - For image features: keyframe-based $\rightarrow$ image classification problems
  - For motion features: video-based approach

Problem with the Video-based Approach



- MED data is noisy $\rightarrow$ the clues to determine an event may appear within a small segment of the entire video.

Our Segment-based Approach



- The basic idea is to examine shorter segments instead of using the entire video.

Comparison

Table: Comparison of different segment-based approaches with the video-based approach on the MED 2011 dataset.

Event/MAP	Non-overlapping sampling			Overlapping sampling			Video-based
	Best (at 90 s)	Late fusion (all lengths)	Late fusion (60, 90, 120 s)	Best (at 120 s)	Late fusion (all lengths)	Late fusion (60, 90, 120 s)	
E006	**0.1277**	0.1217	0.1244	0.1151	0.1086	0.1083	0.0959
E007	0.1521	0.1419	0.1369	0.1552	0.1610	**0.1616**	0.1303
E008	0.4923	**0.4975**	0.4973	0.4969	0.4903	0.4871	0.4766
E009	0.2072	0.2145	0.2064	**0.2160**	0.1954	0.1958	0.0943
E010	0.0916	0.0771	0.0753	0.1008	0.1108	**0.1109**	0.1020
E011	0.0698	0.0805	0.0813	**0.1591**	0.0819	0.0845	0.0609
E012	**0.3560**	0.3309	0.3277	0.3150	0.3293	0.3341	0.2858
E013	0.6030	0.6033	0.6096	**0.6188**	0.5872	0.5910	0.5385
E014	0.2008	0.2585	0.2579	**0.2744**	0.2706	0.2694	0.2138
E015	0.1599	0.1583	0.1622	0.1562	0.1795	**0.1795**	0.0964
All	0.2460	0.2484	0.2479	**0.2607**	0.2515	0.2522	0.2095

Problem with the Video-based Approach

Example video for "assembling a shelter" event in the TRECVID MED 2010 dataset. The top row shows the relevant frames while the bottom row shows the noisy frames.

$\rightarrow$ Handle in the feature aggregation point of view

Traditional Approach

- Building a codebook using K-means, L2 distance, 4000 clusters
- Assign local descriptors to each cluster based on L2 distance
- Features that are assigned to a codeword are pooled to get a representative value for that codeword.

- Traditionally, there are two ways:
  - Sum Pooling: takes a sum over responses to a visual word
  - Max Pooling: select the largest value between feature responding to a visual word

Traditional Approach

- Sum pooling and max pooling techniques can be easily adopted for video representation
- State of art performance with the sum pooling technique in simple video classification/recognition tasks such as sports action videos and studio setting movies

- However, this observation is not true on complex video datasets where the discriminative features may exist within a small part of the video

Sum-max Video Pooling

Toy example:

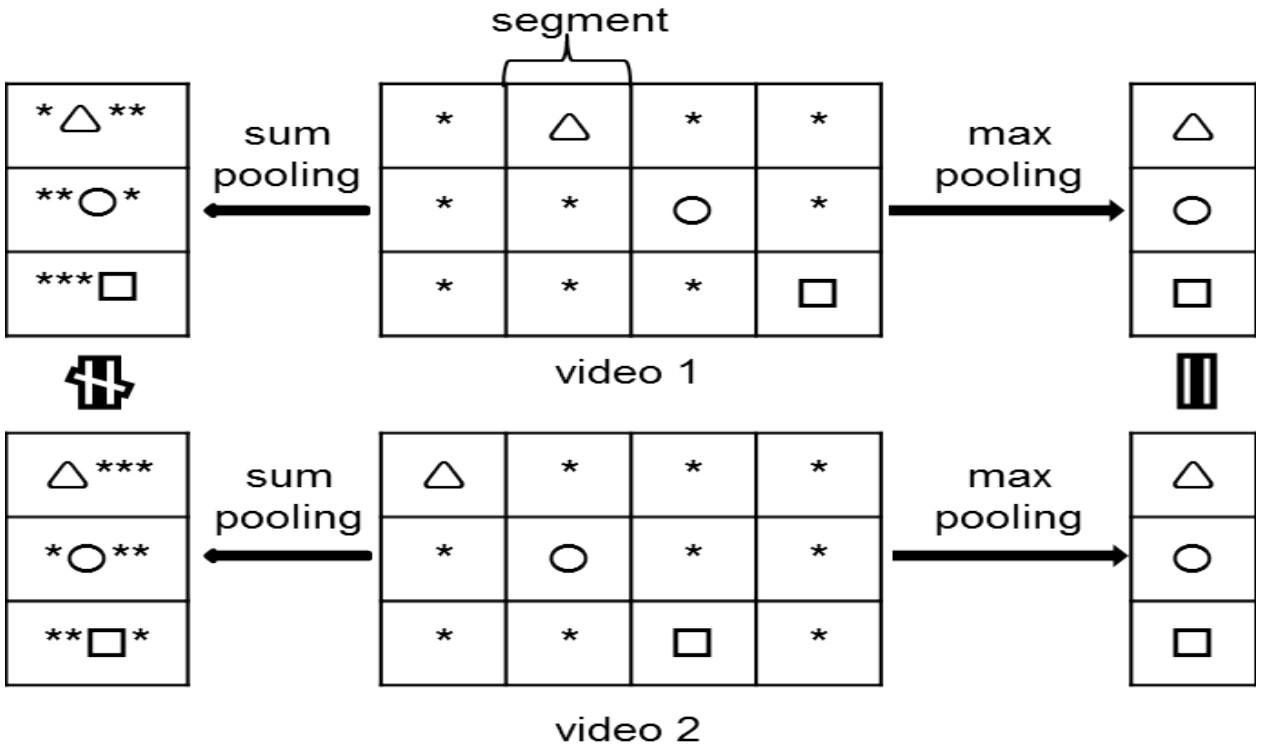
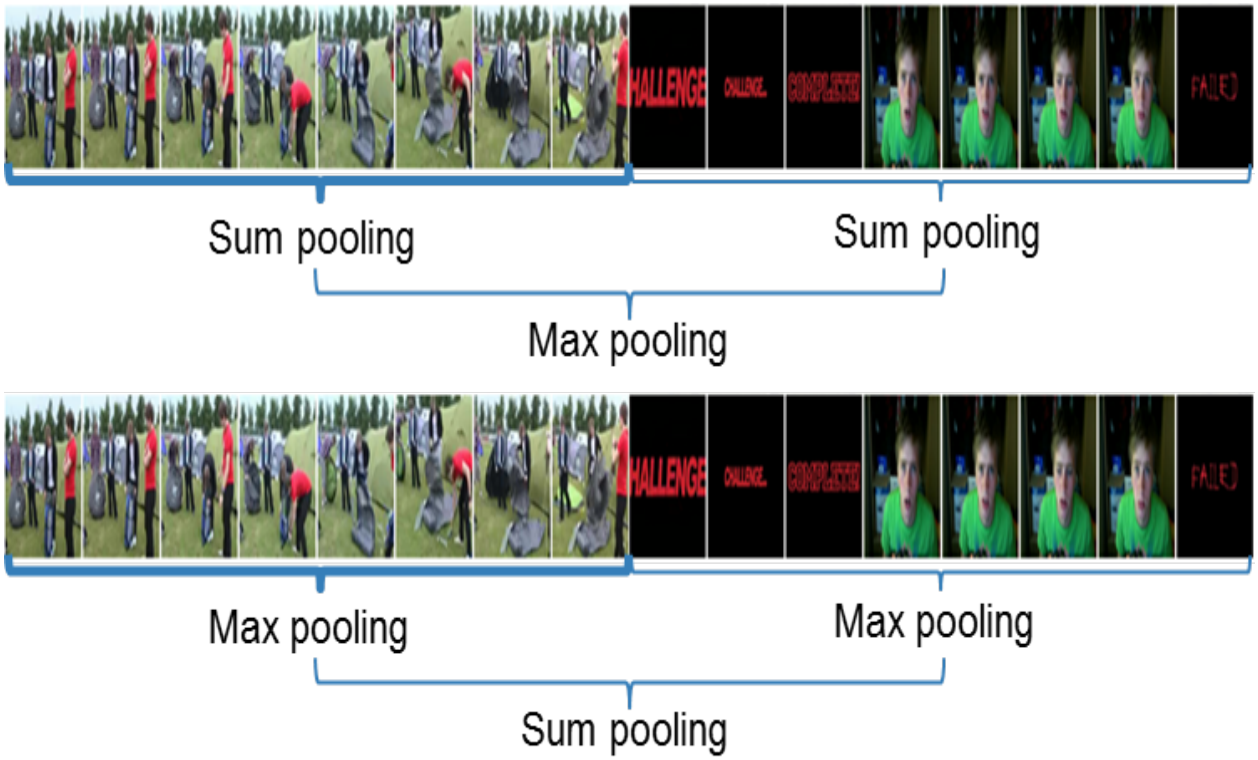


Illustration of sum-max video pooling. △, O, □ represent relevant information; * represents different kinds of irrelevant information, which is popular in complex event data.

Sum-max Video Pooling

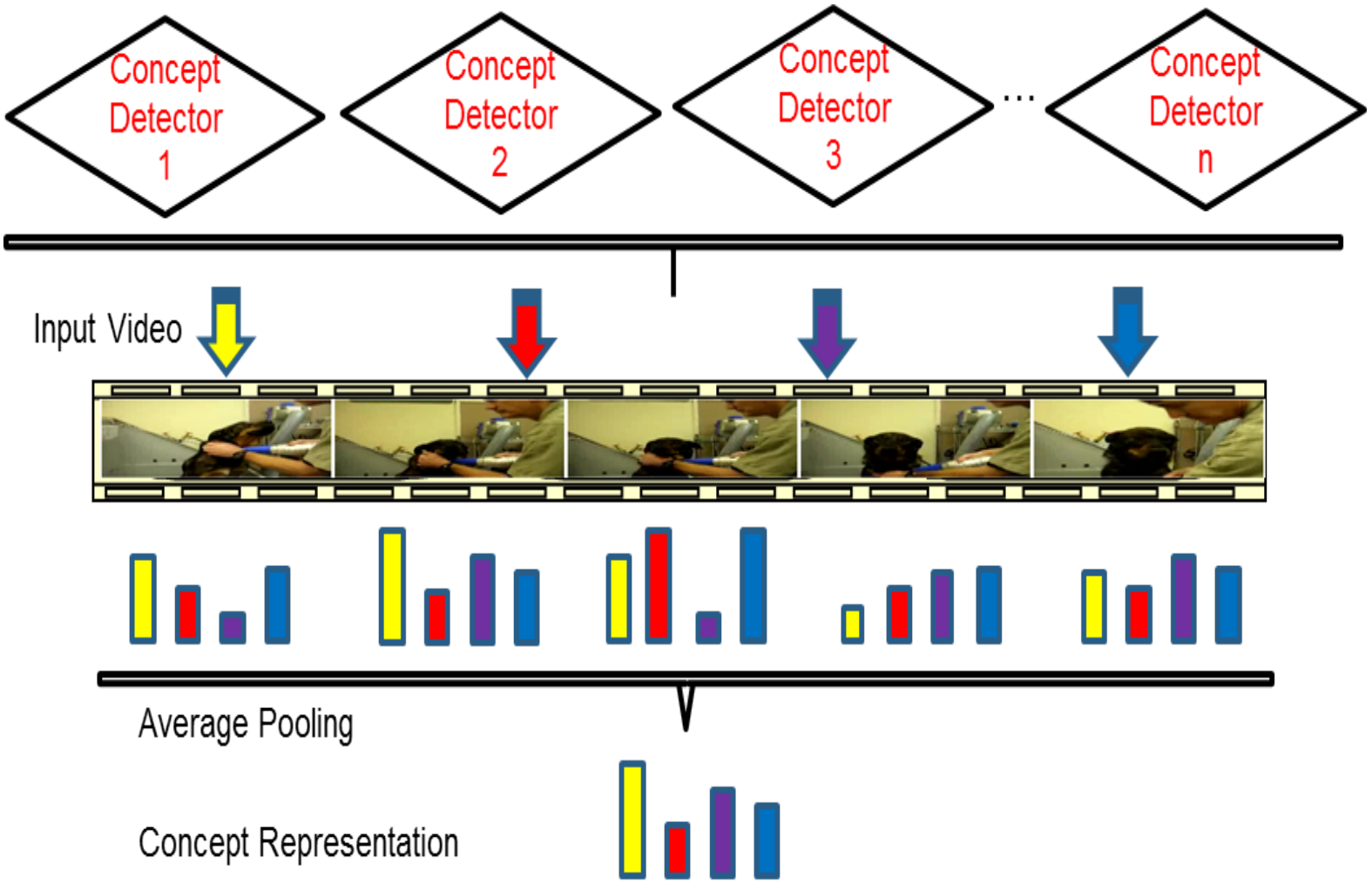


Example of applying sum-max video pooling (top) and max-sum video pooling (bottom) methods on an "assembling a shelter" event video. It can be seen from the top image that after applying max pooling at the segment level, only relevant frames are encoded in the final representation.

Research Questions

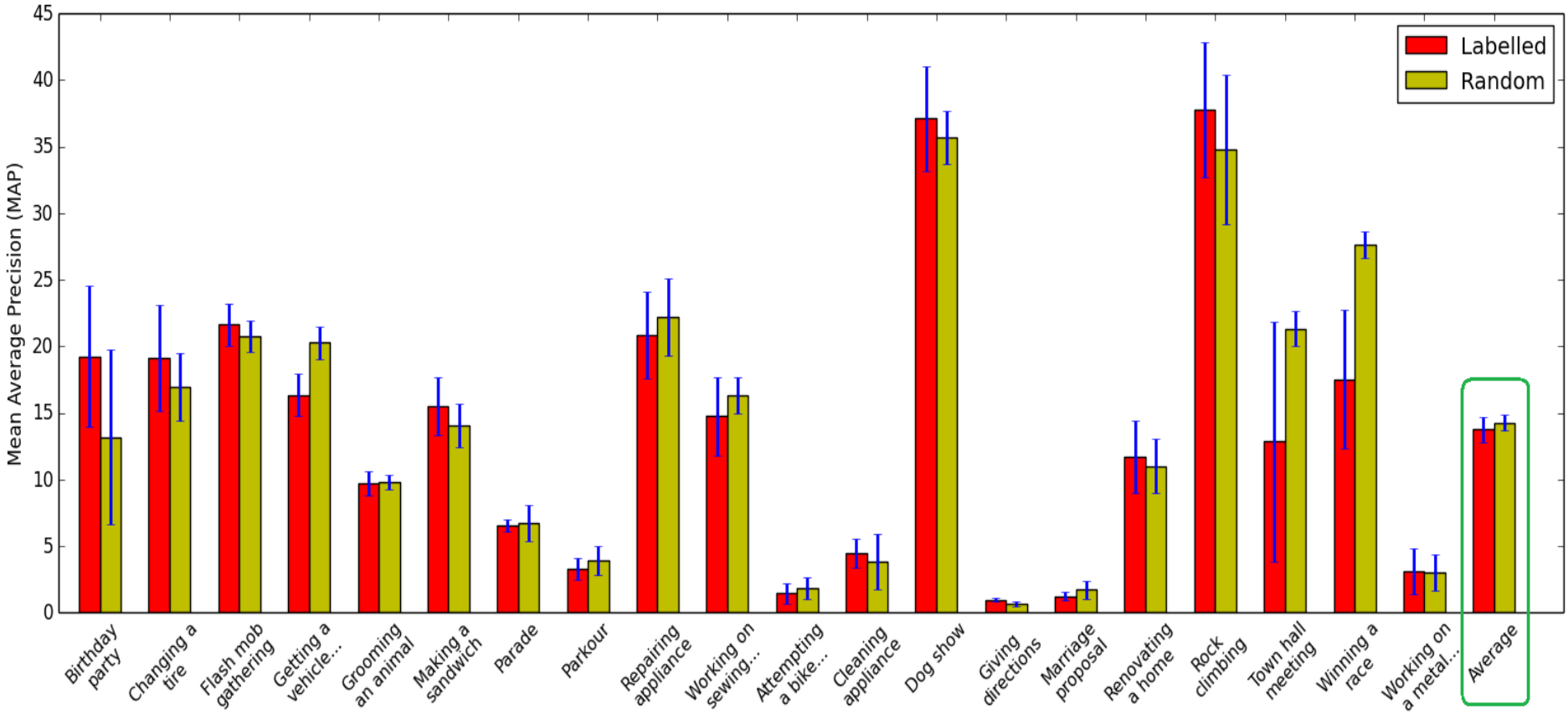
- Labelled vs. Random Concept Detector?
- How Many Random Concepts?
- Influence of Number of Pseudo Positives?

Framework



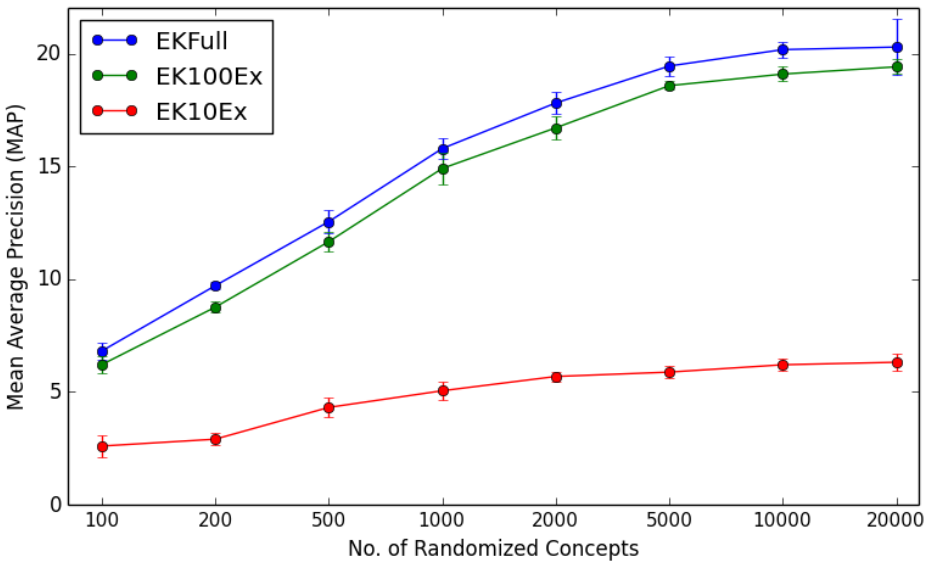
Framework to generate concept-based representation for video.

Labelled vs. Random Concept Detector?

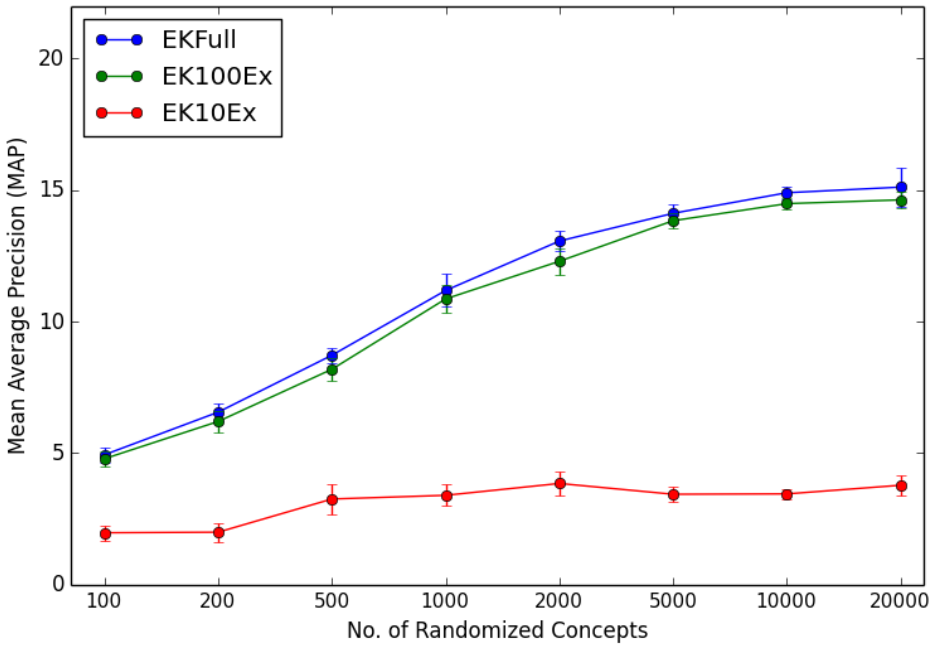


Performance of 1000 labelled concept detectors vs. Performance of 1000 random concept detectors.

How Many Random Concepts?



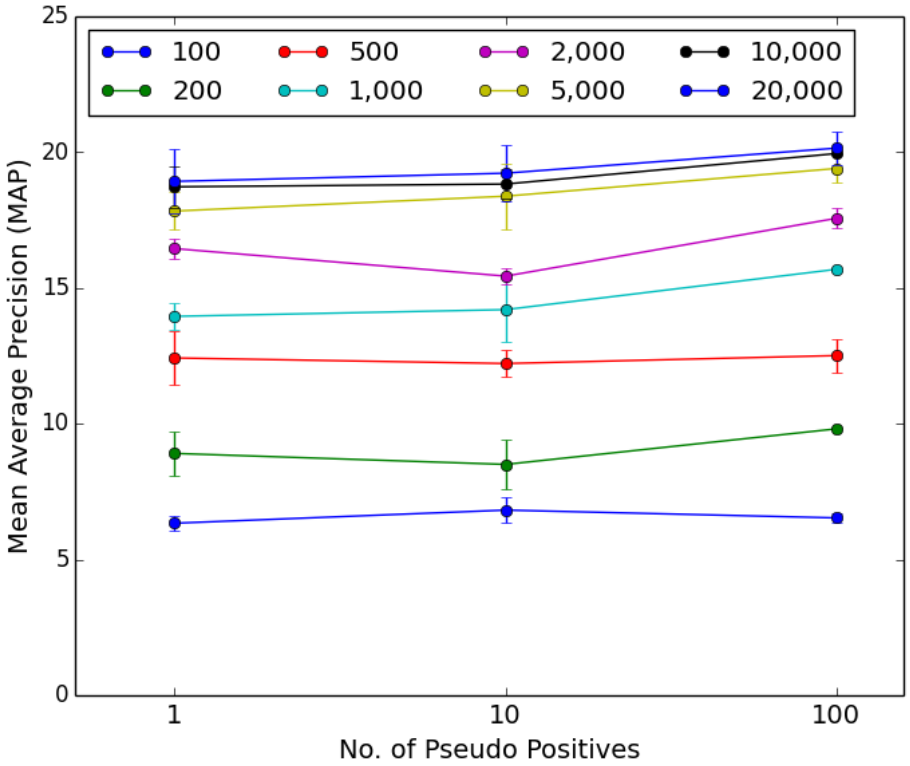
(a) On the MED 2012 KINDREDTEST dataset



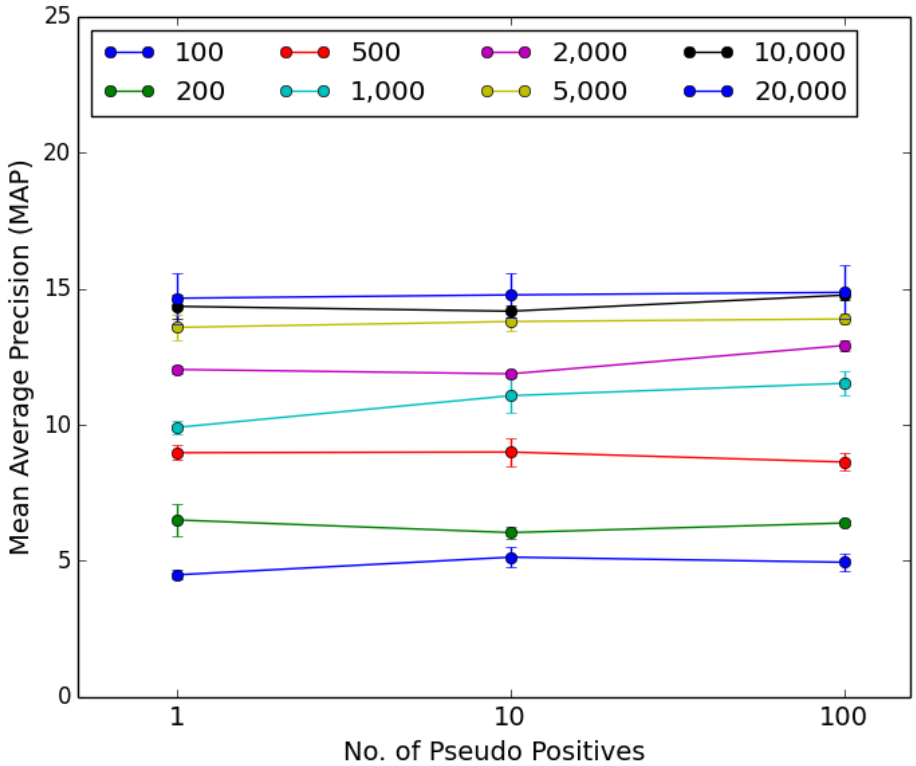
(b) On the MED 2012 MEDTEST dataset

The best performance can be obtained using around ten thousands of totally random labeled concept detectors!

Influence of Number of Pseudo Positives?



(a) On the MED 2012 KINDREDTEST dataset



(b) On the MED 2012 MEDTEST dataset

The detection performance does not depend on the number of pseudo positives!

Issues

- Can we identify key video segments?
- How about classes of video segments? Can be a set of generic classes (generic sub events), or depending of each event (sub events specific to each event)?
- How to combine the analysis results of video segment?
- How to adapt to ad hoc case?
  - key video segments should be automatically detected
  - generic set of sub events should be pre-defined
  - or sub events should be adaptively generated from given event kit