Multimedia Event Detection Using Segment-Based Representation

Sang Phan

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

Dec 17th, 2014

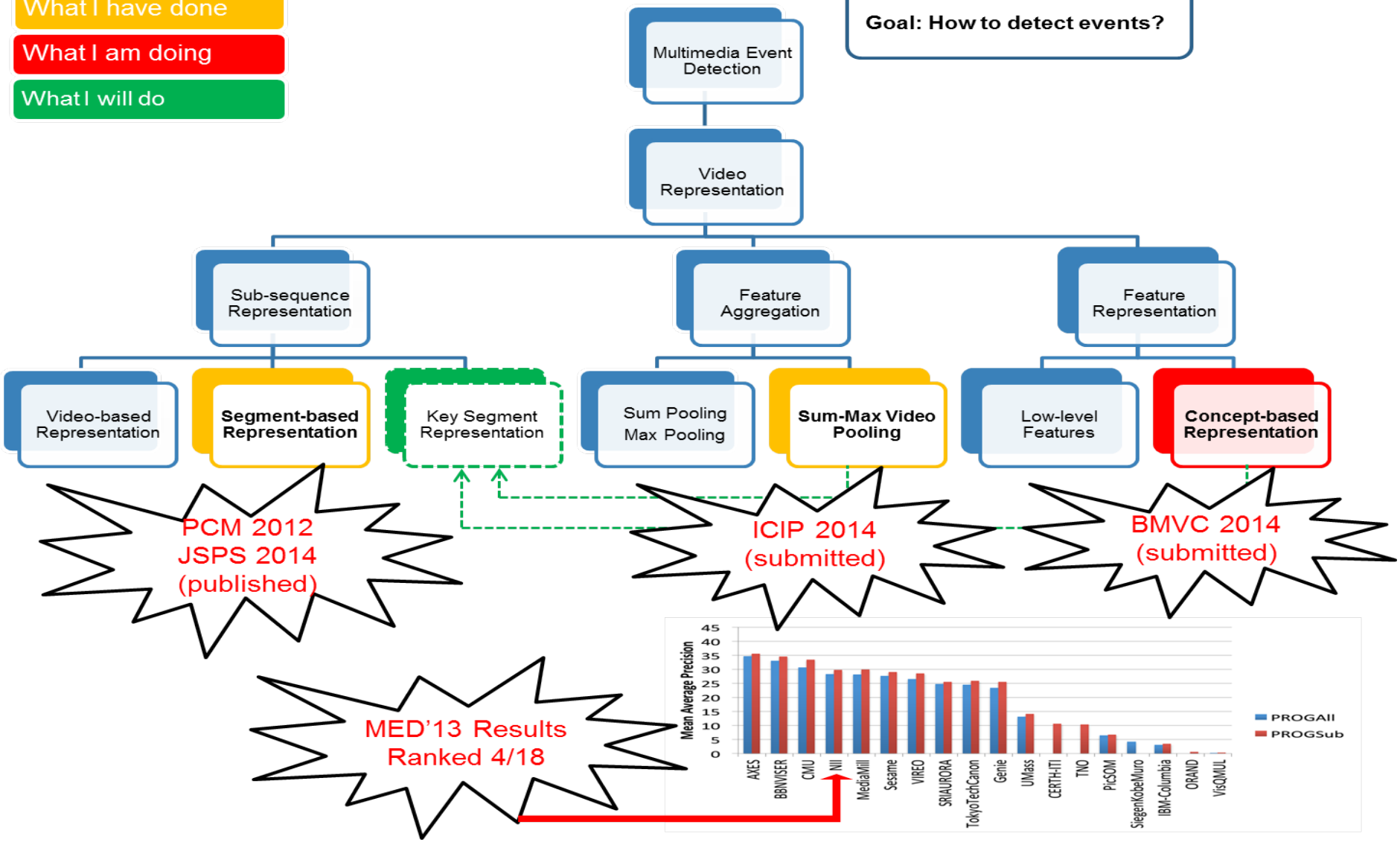
Summary

What I have done

What I am doing

What I will do

Goal: How to detect events?



What happened?

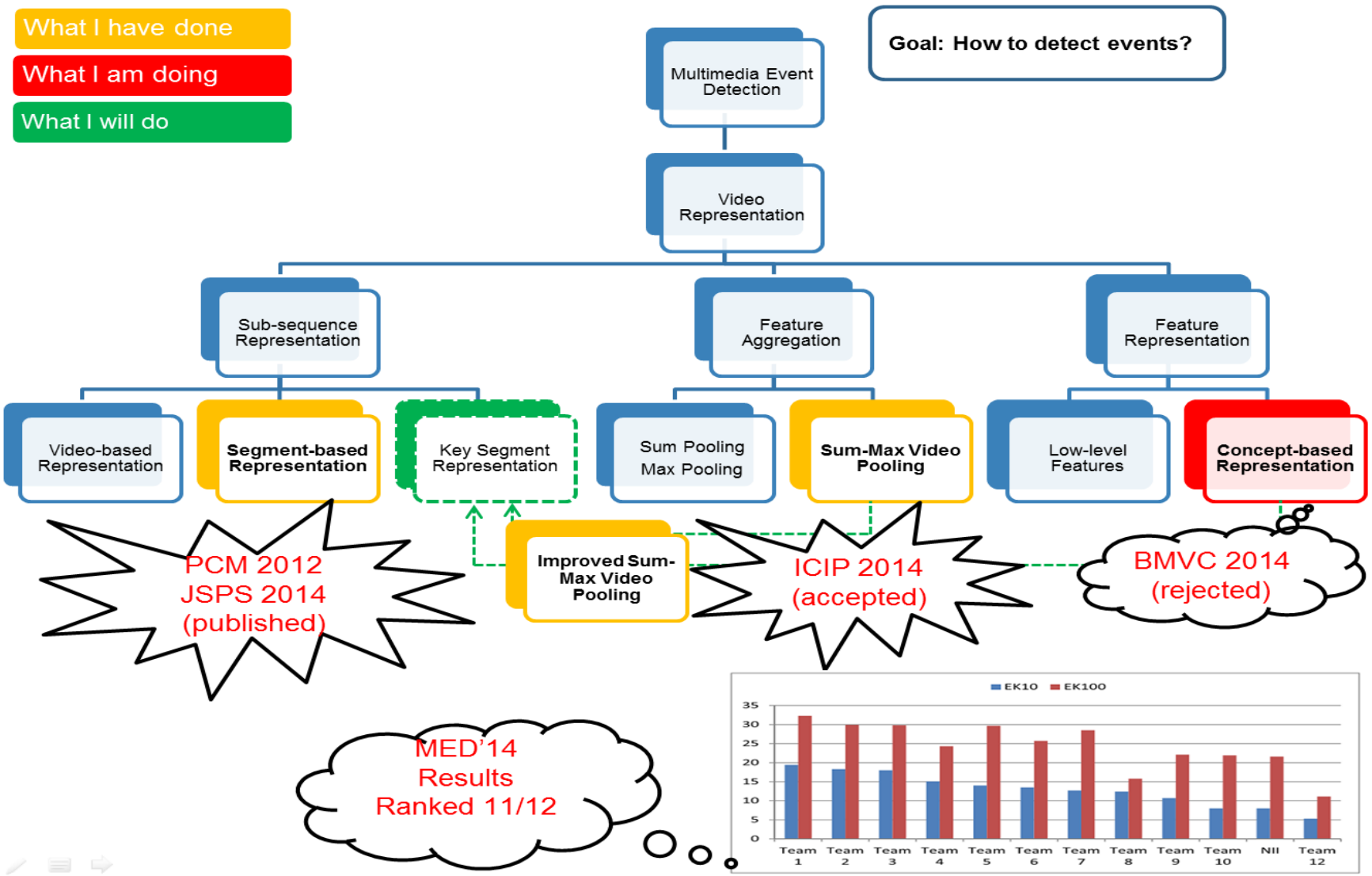


Table of Contents

Table of Contents

Multimedia Event Detection

Motivation

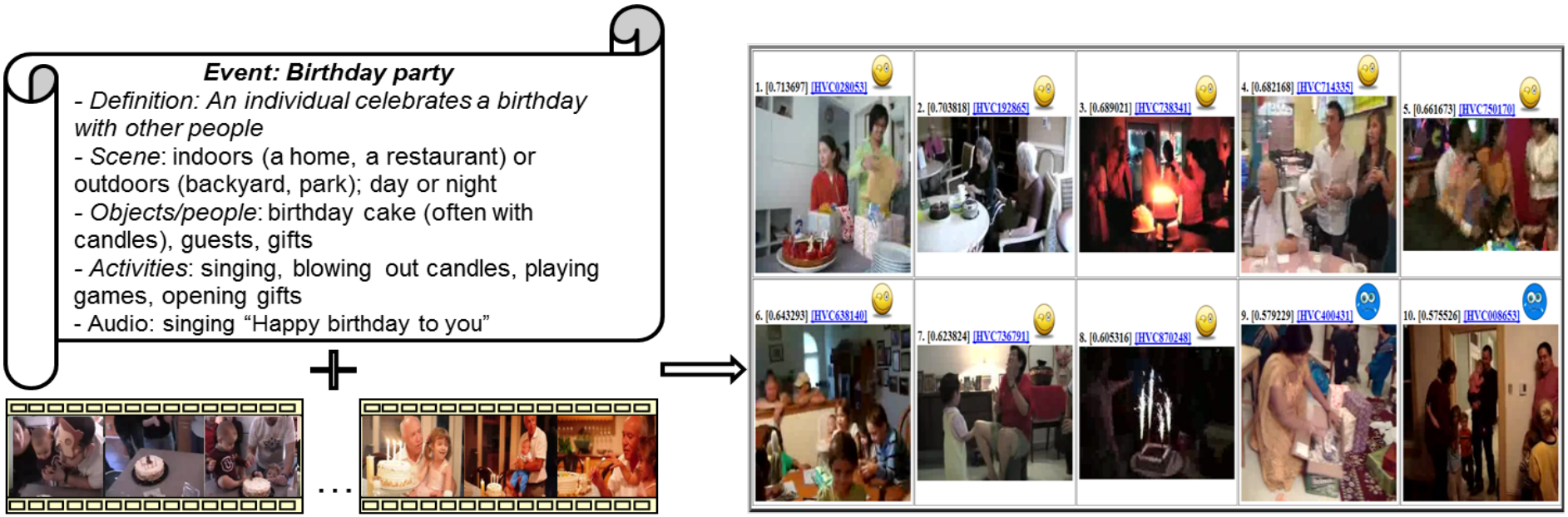
- Massive number of videos are produced every day.
  - YouTube: 72 hours uploaded per minute, with 3 billion viewers a day.
- Video need to be indexed, searched based on its content.
- Many applications:
  - User demands: tutorial videos such as "**how to make a cake**", "**how to repair an appliance**".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

Multimedia Event Detection

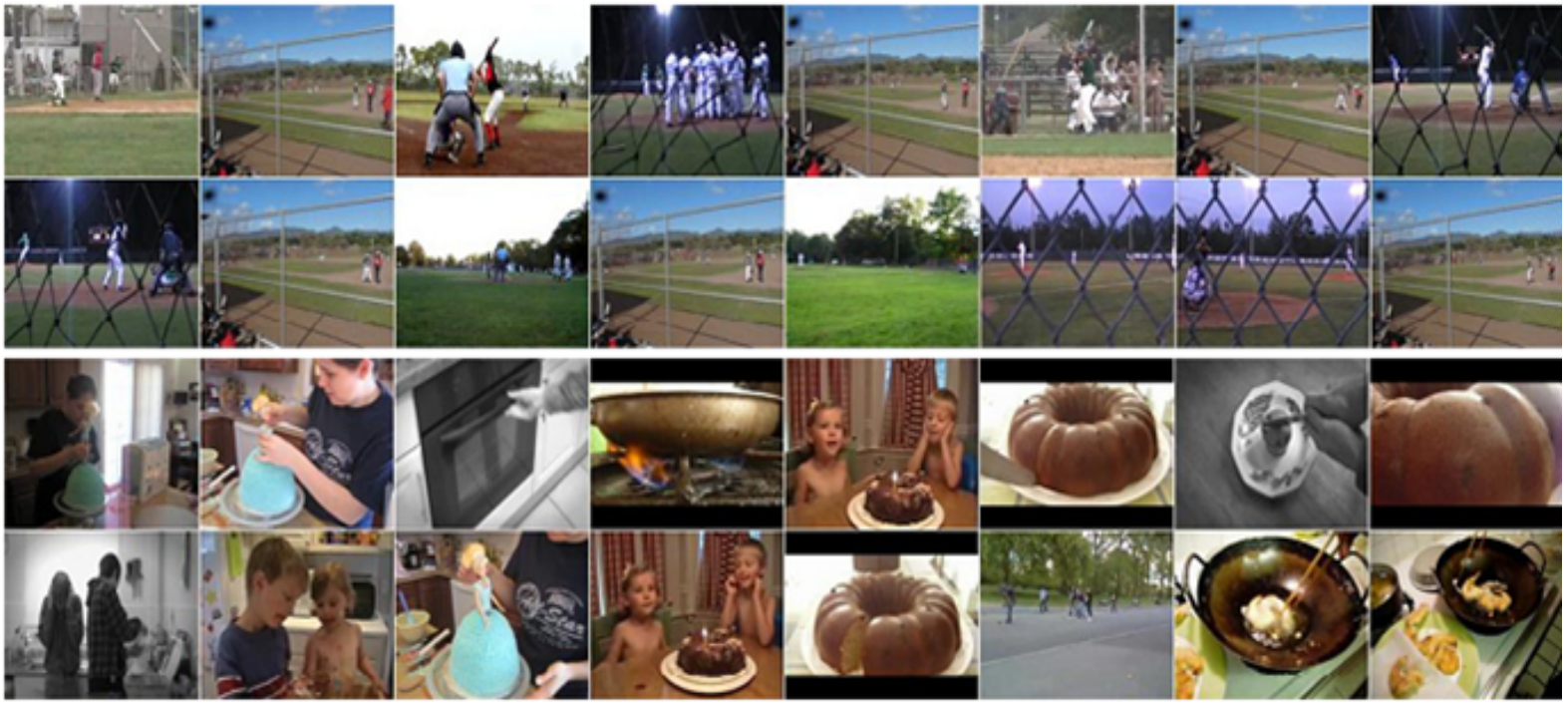
Task defined by TRECVID since 2010

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.

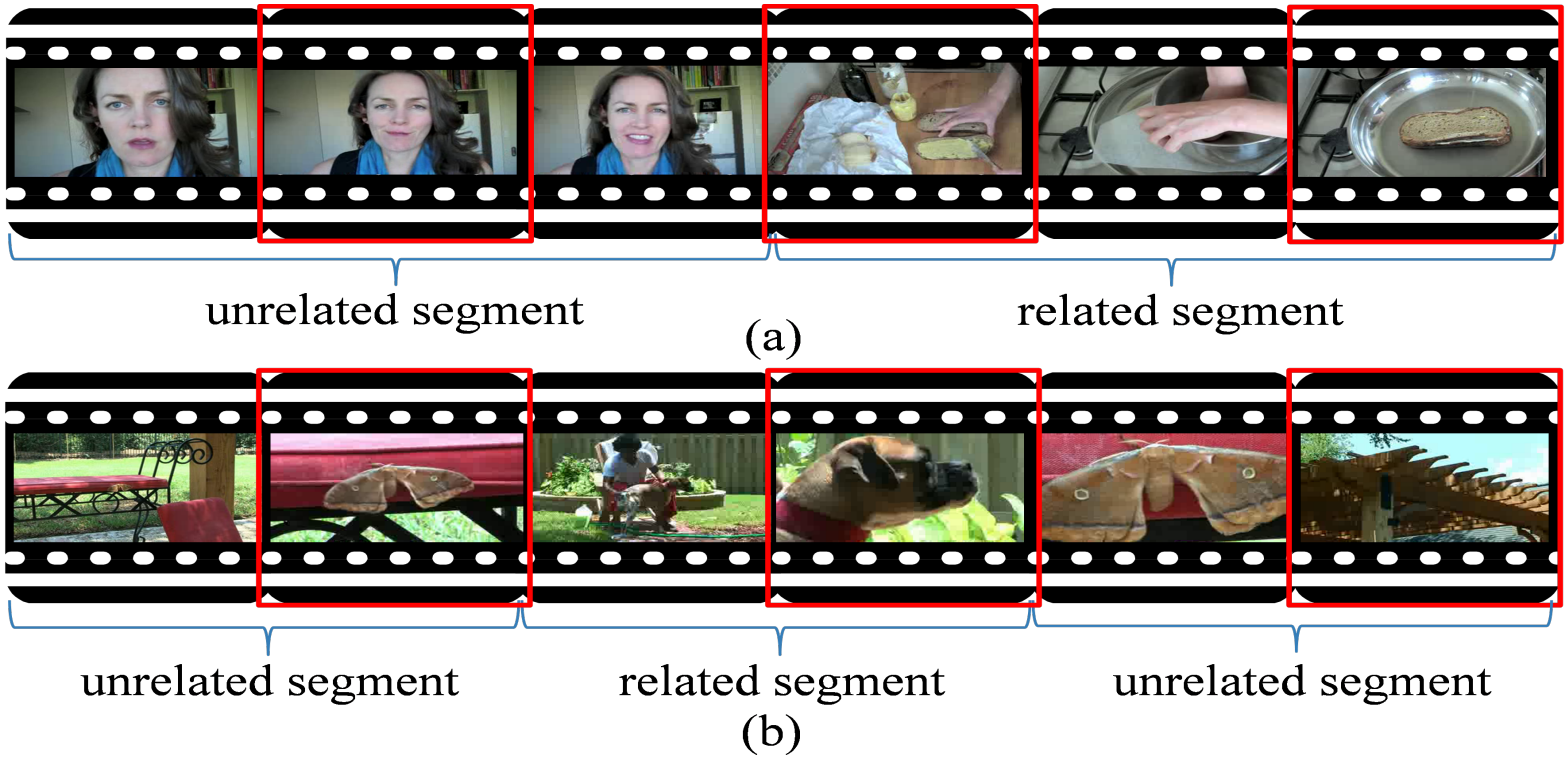
Challenges of Multimedia Event Detection

- 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

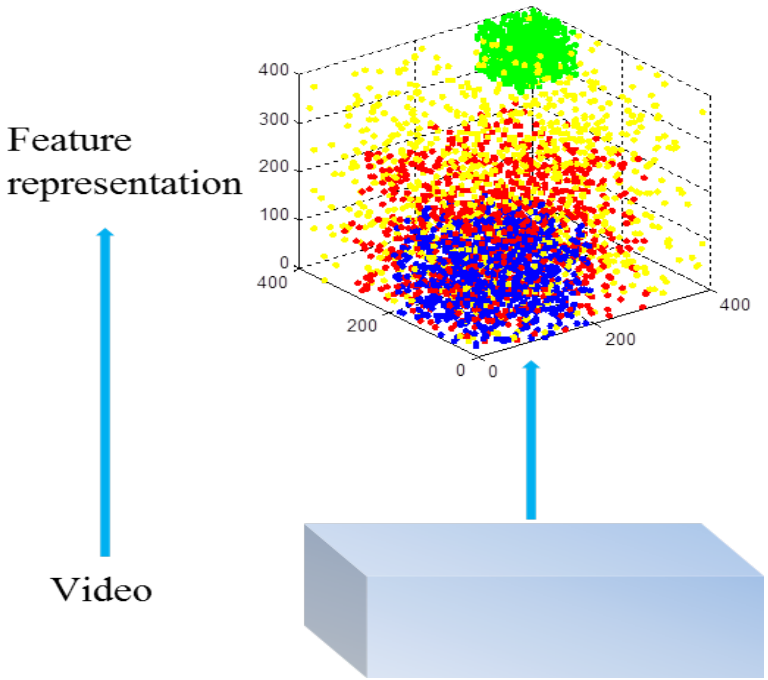


(a) Example video for "making a sandwich" event: the related segment appears after a self-cam segment (unrelated); (b) example video for "grooming an animal" event: related segment is sandwiched between two unrelated segments. This kind of video is popular in realistic dataset.

Table of Contents

Video-based Approach

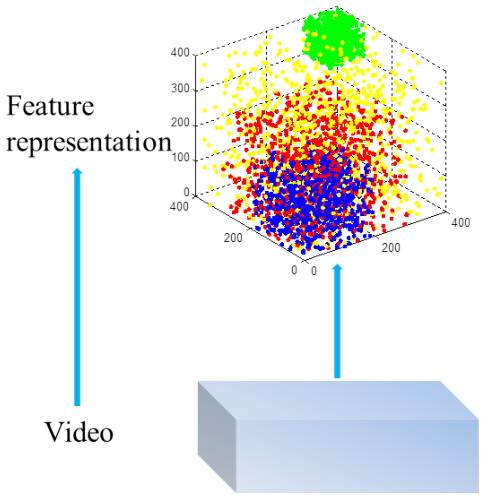
- Features are computed over the whole video
- One representation for each video



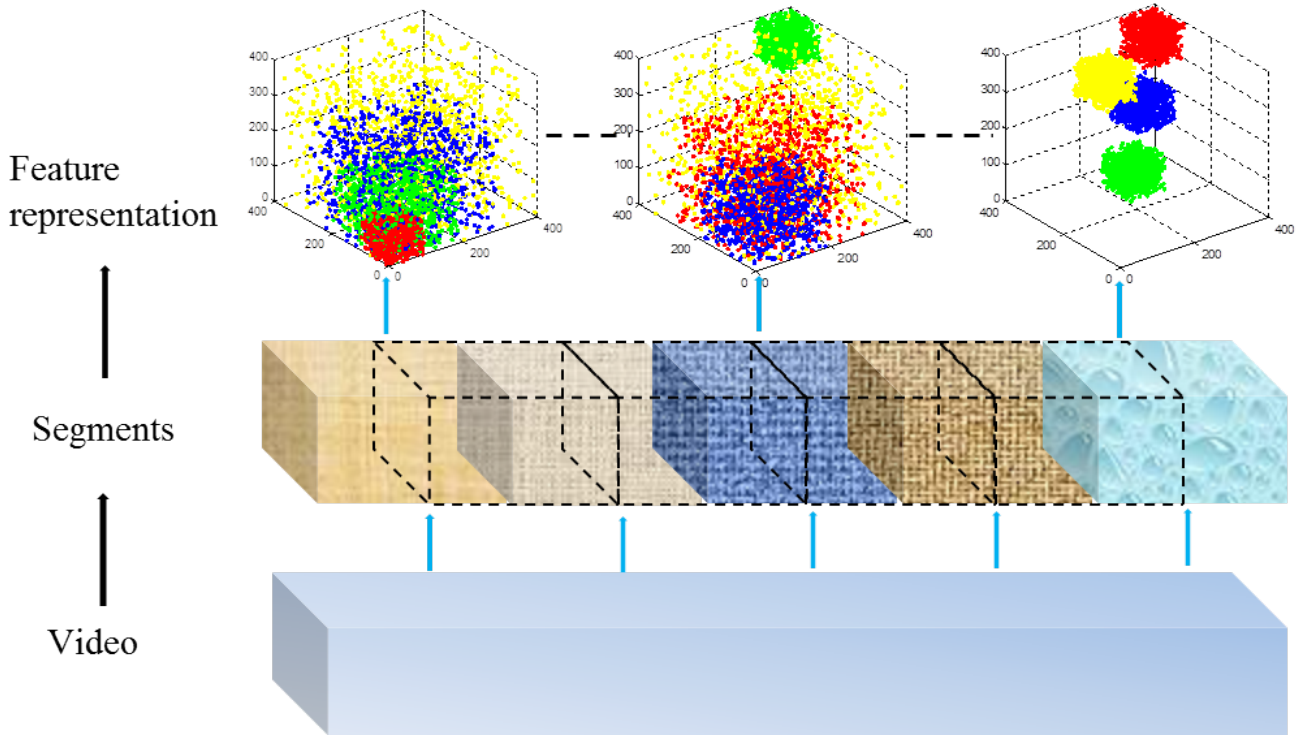
- Used by best MED'10 system (Columbia University)
- Used by best MED'11 system (BBN VISER)

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

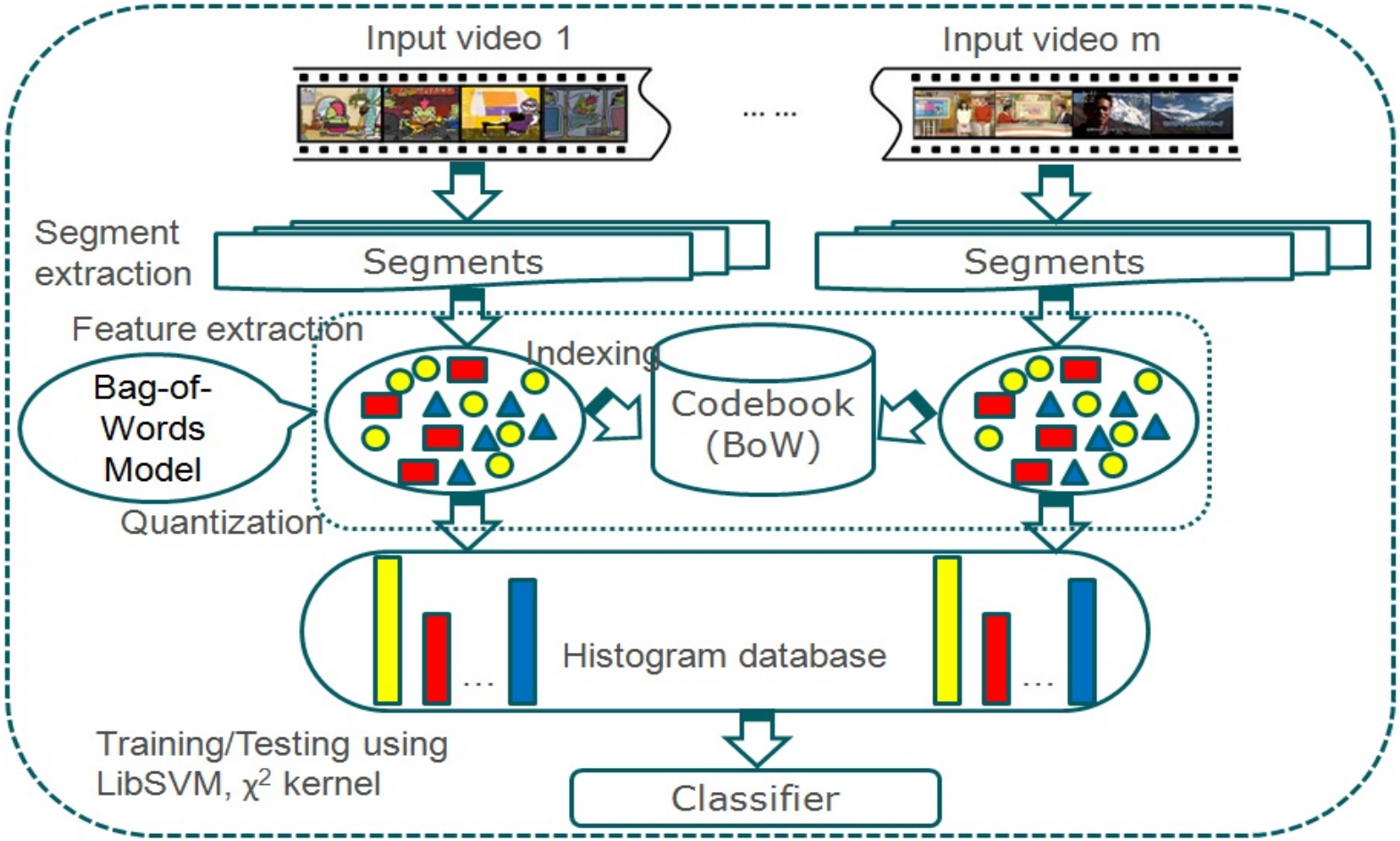
Our Segment-based Approach

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]

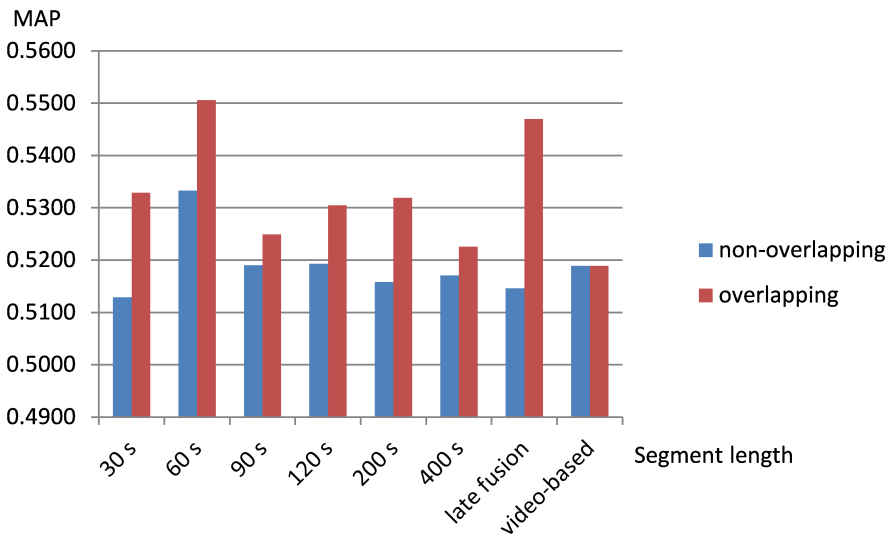
Guimaraes, S.J.F., Couprie, M., Araujo, A.d.A., Leite, N.J: Video segmentation based on 2d image analysis. Pattern Recognition Letters, 2003, 24(7), 947957.

Evaluation Framework

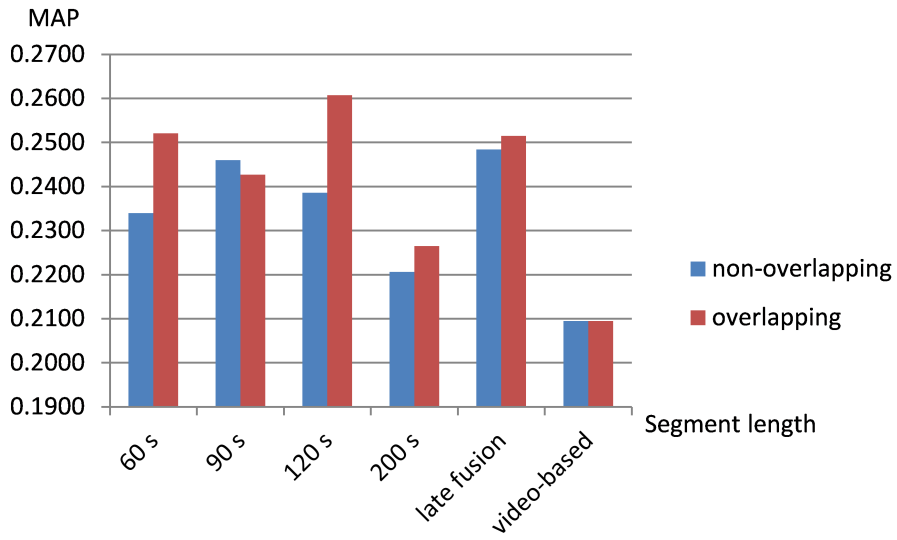


Evaluation framework for our baseline MED system

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.

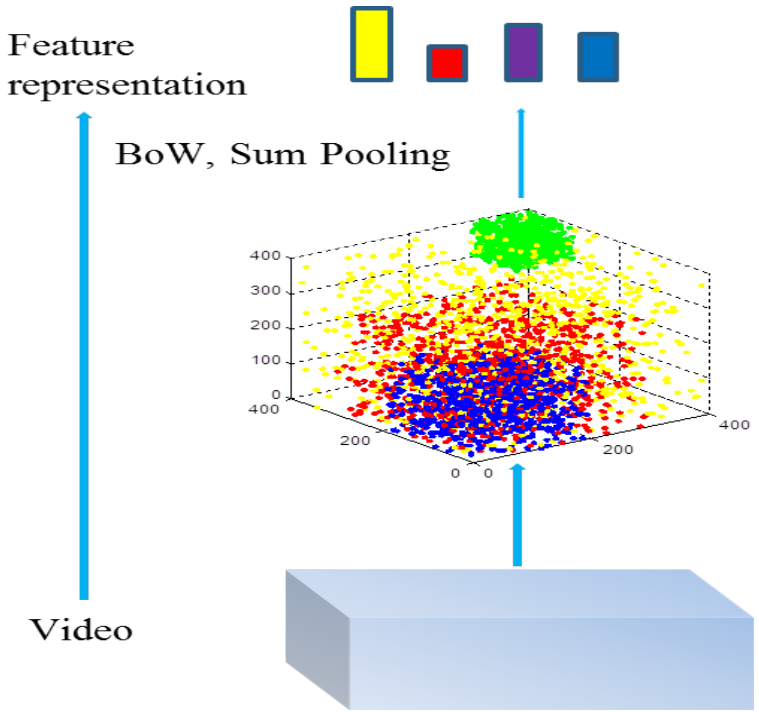
Event/MAP	Best non-overlapping	Best over-lapping	SBD seg-ments	Video-based
Assembling shelter	0.4511	0.4781	0.4284	**0.4911**
Batting in a run	0.7852	**0.7918**	0.7866	0.7902
Making a cake	0.3636	**0.3819**	0.1918	0.2755
All	0.5333	**0.5506**	0.4689	0.5189

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

Table of Contents

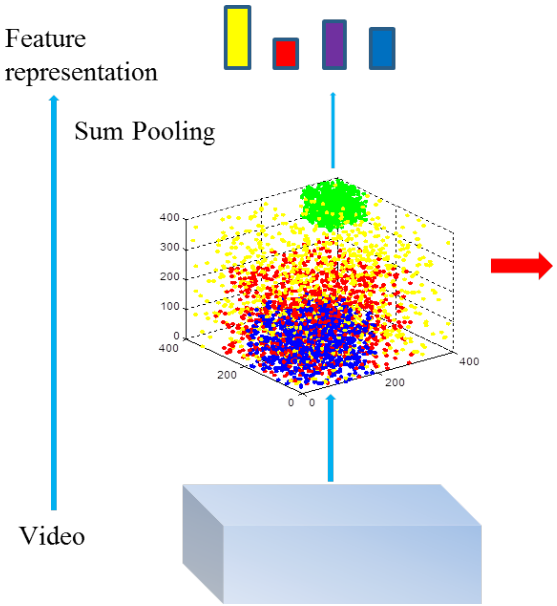
Video-based Approach

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

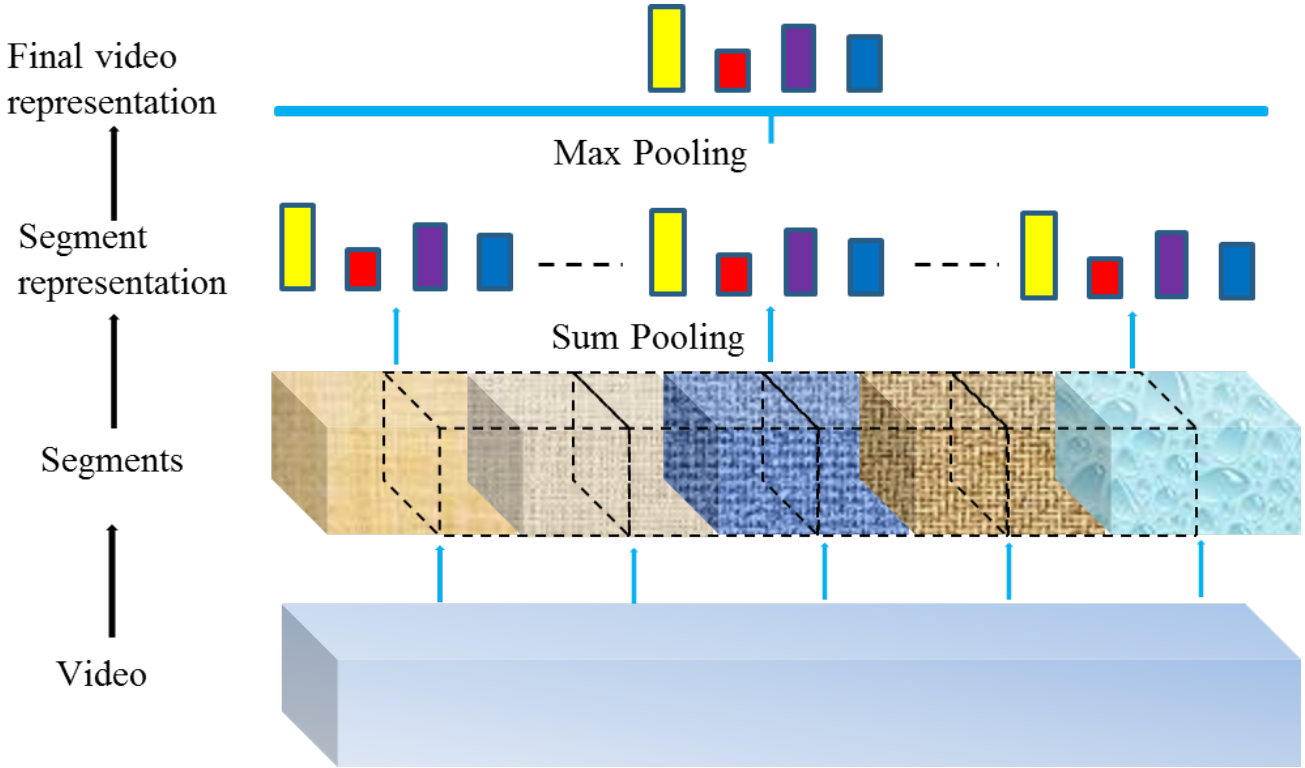


- The proposed segment-based approach: increasing the number of segment representation is not scalable.
- **Specific problem:** How to generate one representation per each video from its segment-level representations?

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:

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

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

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?

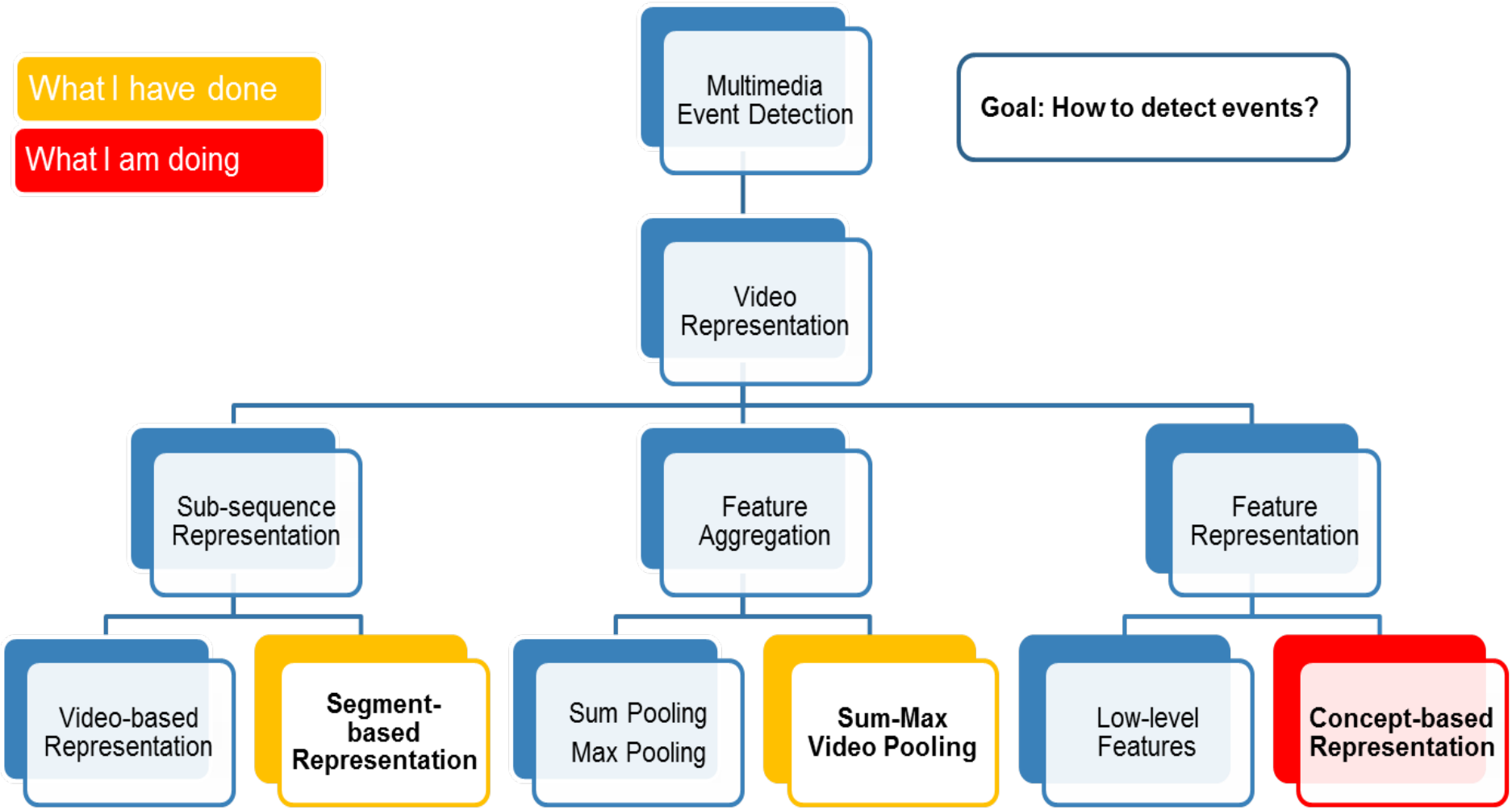


Table of Contents

Limitation of Uniform Segment Sampling

- Segment too short: relevant segments might span to several segments.

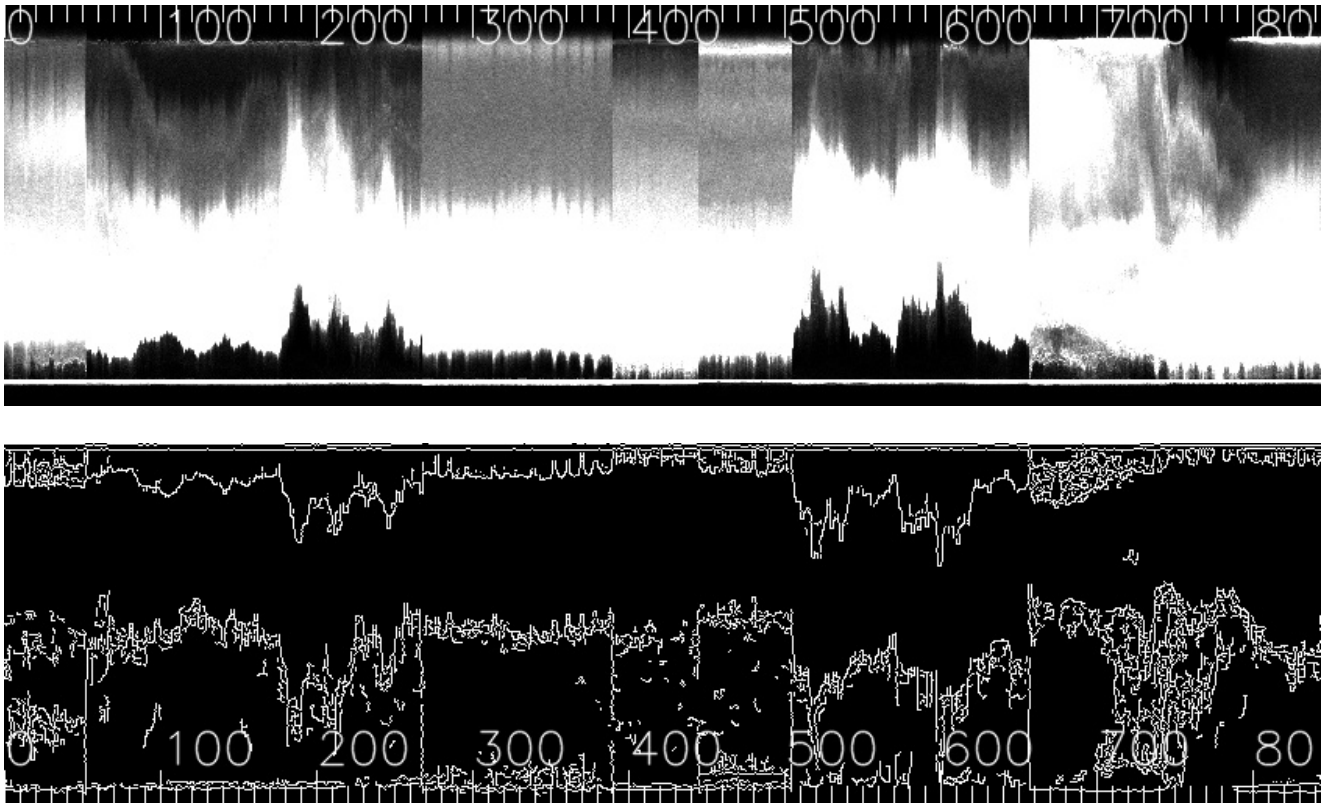
- Segment too long: can contain irrelevant information.

- It is also difficult to find the optimal segment length.

Typical Shot Boundary Detection

Based on the global similarity between two consecutive frames.

- Shot Boundary Detection [Guimaraes et al. - 2003].

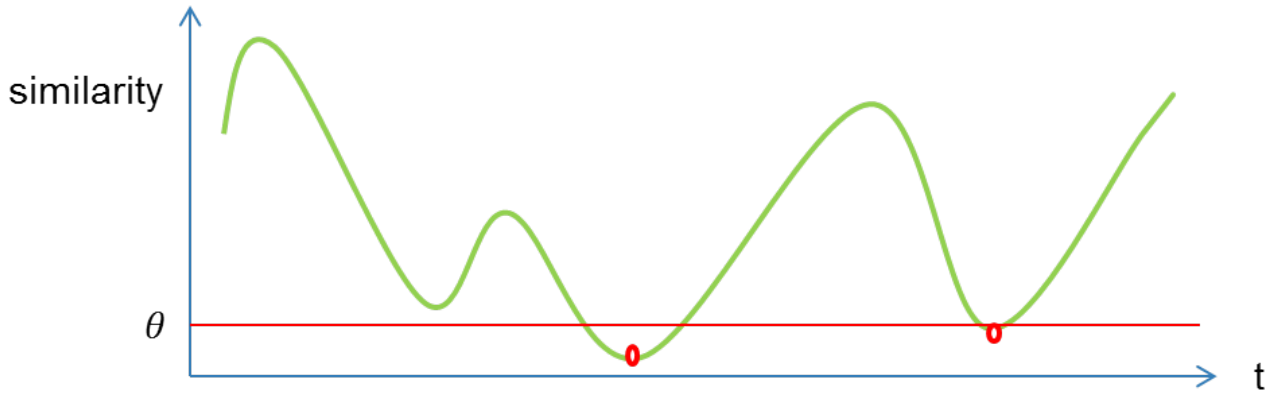


Guimaraes, S.J.F., Couprie, M., Araujo, A.d.A., Leite, N.J: Video segmentation based on 2d image analysis. Pattern Recognition Letters, 2003, 24(7), 947957.

Improved Shot Boundary Detection

Based on the semantic similarity between two consecutive keyframes.

- Extract sample keyframes at a fix intervals (eg. 2 seconds)
- Extract semantic features for each keyframes using DeepCaffe (trained on ImageNet1K) [Jia et al., 2014]
- Calculate the similarity between two consecutive keyframes (eg. Cosine similarity)
- Calculate local minimas and apply thresholding



Jia et al.:Caffe: Convolutional Architecture for Fast Feature Embedding. arXiv preprint arXiv:1408.5093, 2014.

Benefits of Improved Shot Boundary Detection

- Better video segmentation method for MED videos

- Robust to illumination changes

- Can provide semantic information, easy to build a semantic hierarchy

Tusker: 0.519
African elephant: 0.150
Indian elephant: 0.126
Bison: 0.053
Ox: 0.031

Results of Improved Sum-max Video Pooling

Same experimental setup with the previous Sum-max video pooling framework.

θ	0.005	0.0083	0.0139	0.0232	0.0387	0.0646	0.1077	0.1797	0.2997	0.500	sum pooling
assembling_shelter	0.479328	0.514231	0.506029	0.512407	0.513717	0.517323	0.515518	0.520276	**0.520098**	0.491107	0.446847
batting_in_run	0.8141	**0.816593**	0.815027	0.779839	0.779954	0.76251	0.768743	0.769599	0.779679	0.761805	0.798751
making_cake	0.341899	0.330765	**0.343695**	0.335386	0.297625	0.283869	0.293041	0.29665	0.304311	0.325143	0.305314
all	0.545109	0.553863	**0.554917**	0.542544	0.530432	0.521234	0.525768	0.528842	0.534696	0.526018	0.516971
segment length	71 s	60 s	51 s	n/a	n/a	n/a	n/a	n/a	n/a	10 s	

- Less sensitive to similarity threshold θ.
- Best performance can be obtained with average segment length around 60 s.

Comparison

Event/MAP	Sum pooling	Sum-max pooling (best at 60s)	Sum-max Video pooling (SBD, 60s)	**Improved Sum-max pooling** (best at 50s)	Columbia U Best of MED10 (STIP)
assembling_shelter	0.446847	**0.507175**	0.4476	0.506029	0.468
batting_in_run	0.798751	0.790048	0.8090	**0.815027**	0.719
making_cake	0.305314	0.309965	0.3097	**0.343695**	0.476
all	0.516971	0.535729	0.5221	**0.554917**	0.554

- Improved Sum-max video pooling outperforms other baselines.
- Using the exact shot boundary is important for video pooling.
  (Note that we compare with the SBD algorithm at the same average segment length)
- Comparable performance with best MED system in TRECVID 2010.

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

Toward Key Segment Selection for MED

Definition

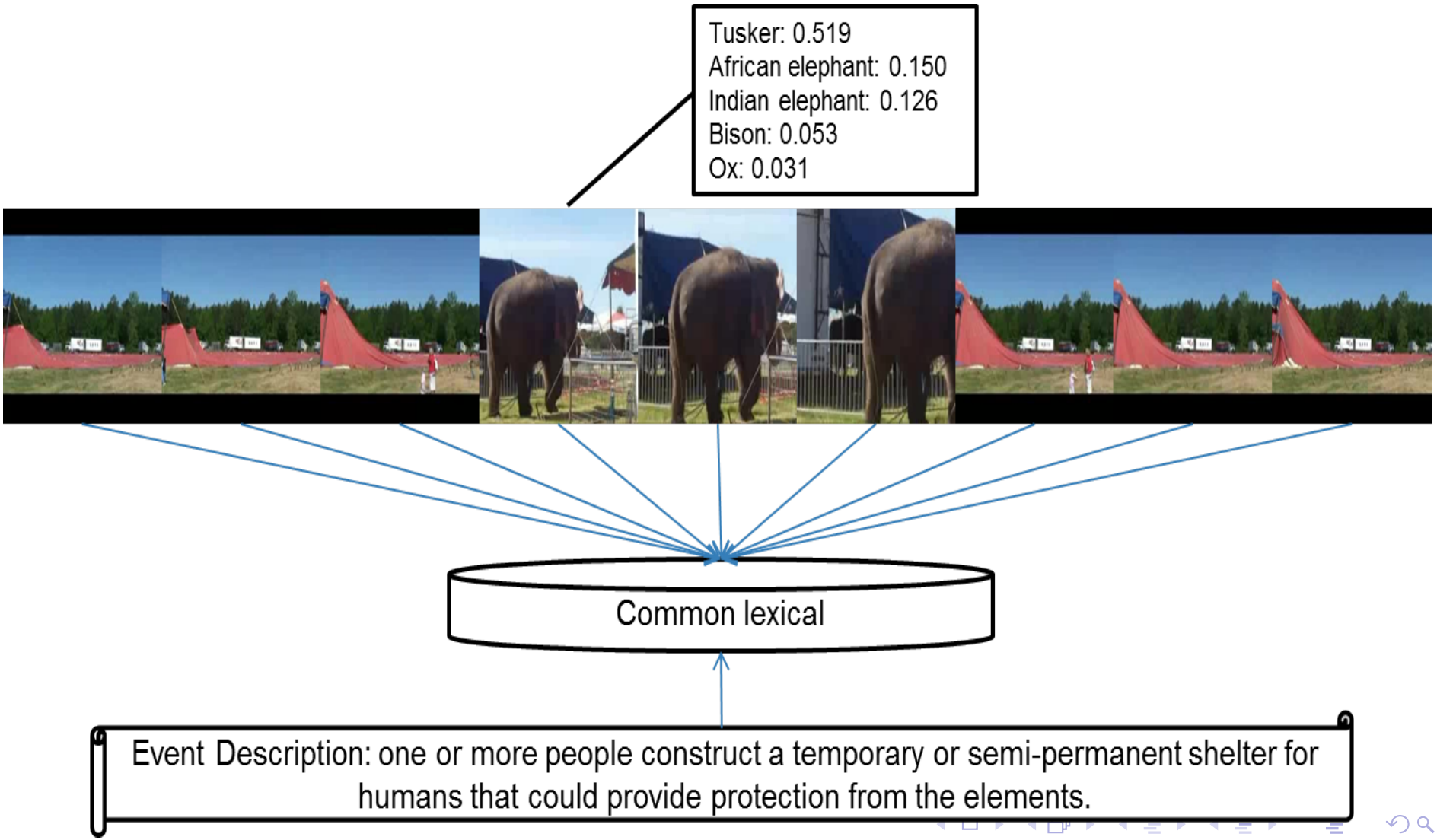
Key segments: segments that contain positive evidence for a specific event

- 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?

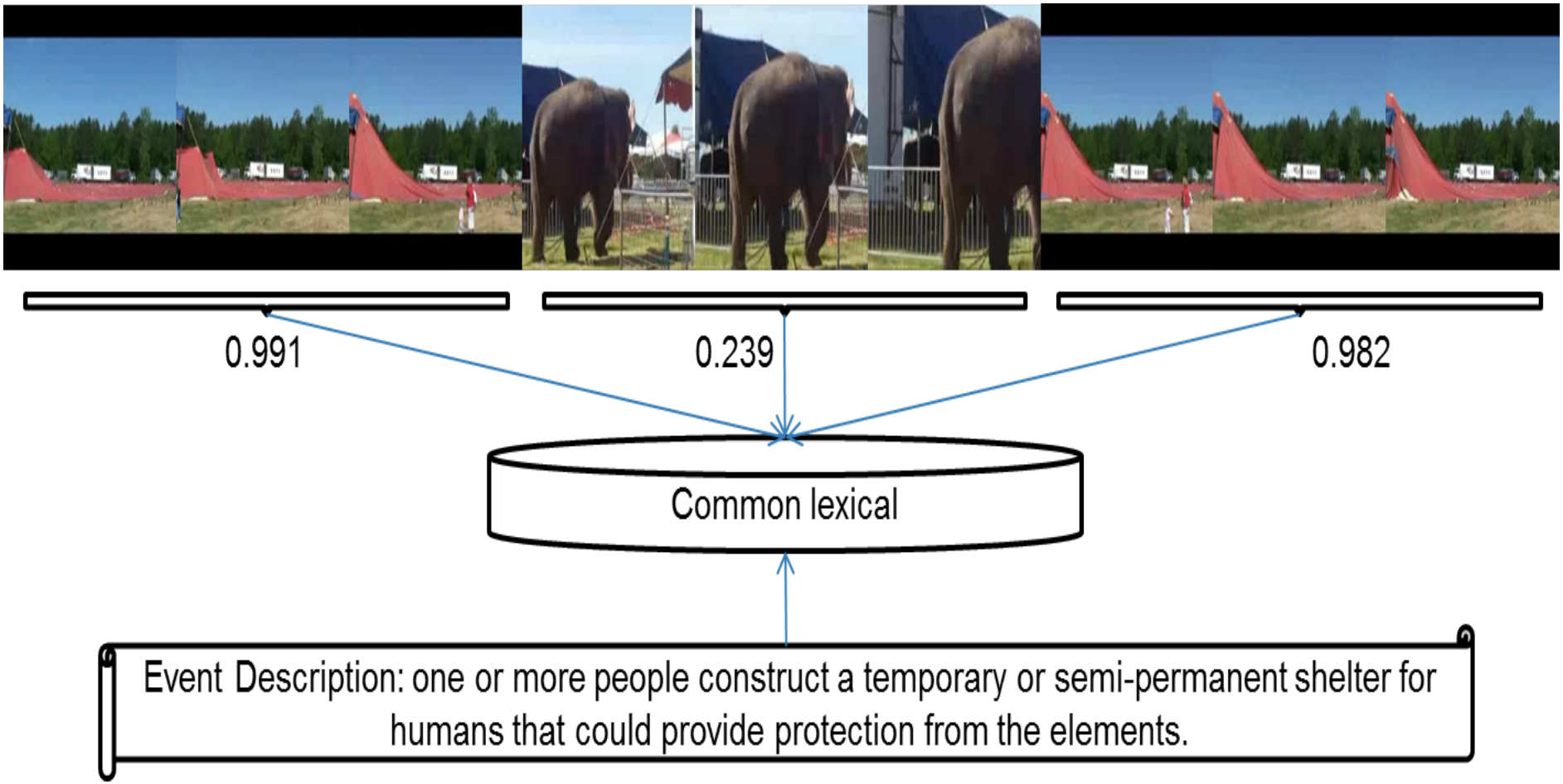
Toward Key Segment Selection for MED

Mapping the results of visual processing and text processing into a common space.



Toward Key Segment Selection for MED

- Merging keyframe similarity to calculate segment similarity.
- Key segments are segments that have high similarity to the event description.



Then...

- Apply Sum-max Video Pooling.
- Extend to Weighted Sum-max Video Pooling.
- How to impose temporal relationship between segments?
  - Assembling a shelter: people should gathering before assembling

  - Birthday party: people often singing before blowing candle and then eating

$\rightarrow$ Language knowledge might help to incorporate these constraints.

Automatic Video2Text Generation

[collaboration work with Prof. Yusuke Miyao, NII]

- We aim to generate text descriptions for key segments, as well as the whole video.

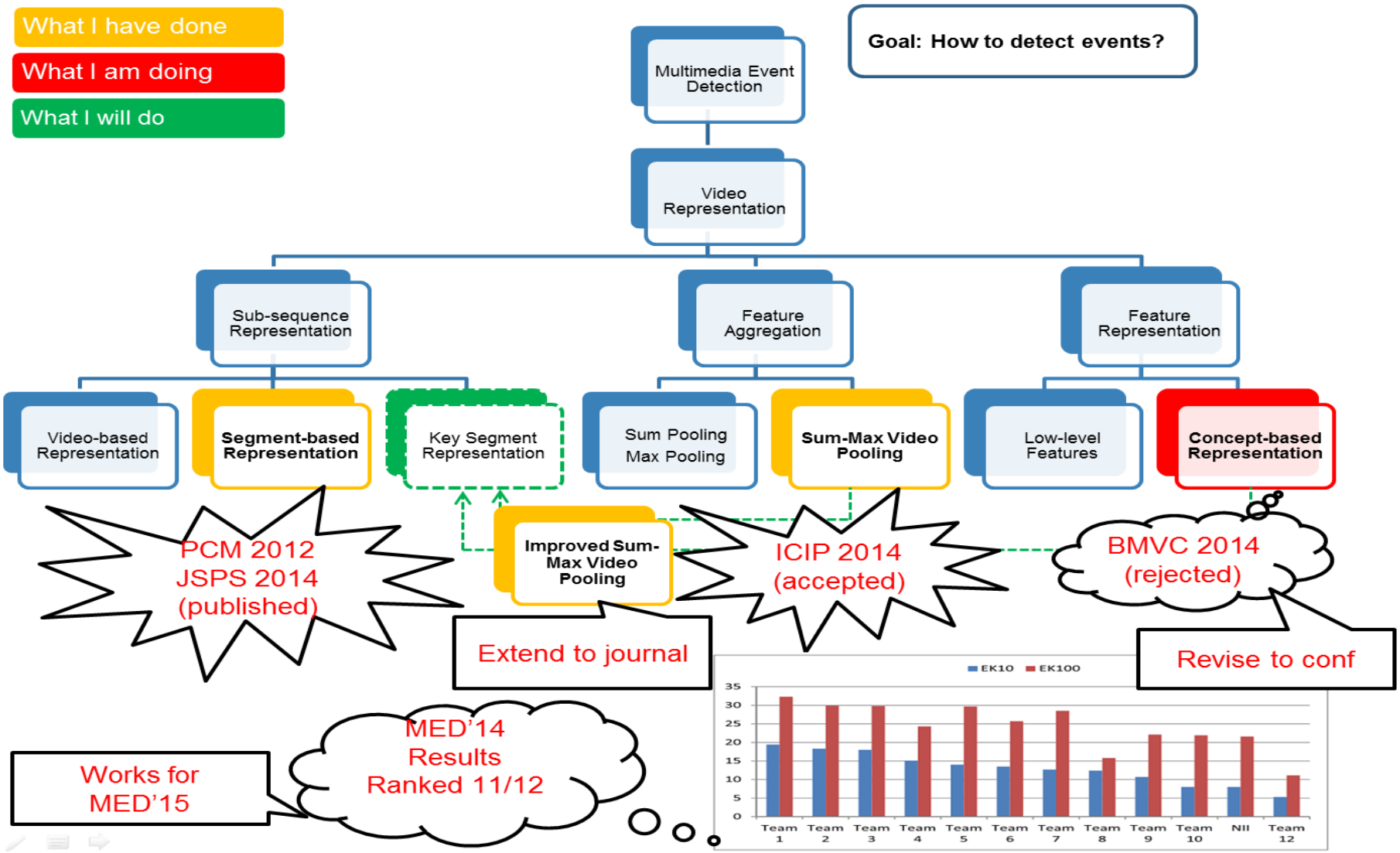
an event for little children where they get to get games and food

a woman swimming in a metallic frame is played on a video screen hanging on a wall.

People singing around a table

- The results of this project can be used for event recounting, i.e. provide evidences for event detection.
- We are building a VideoNet for action classification in real world situation.

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)
Improved Sum-Max Video Pooling
ICIP 2014
(accepted)
BMVC 2014
(rejected)
Extend to journal
Revise to conf
MED'14
Results
Ranked 11/12
Works for
MED'15
EK10
EK100
35
30
25
20
15
10
5
0
Team 1
Team 2
Team 3
Team 4
Team 5
Team 6
Team 7
Team 8
Team 9
Team 10
NII
Team 12

Thank you for your attention!