

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

- 1 Multimedia Event Detection
- 2 Segment-based Representation
- 3 Sum-Max Video Pooling
- 4 Sum-Max Video Pooling with Better Video Segmentation
- 5 Next Study

Table of Contents

- 1 Multimedia Event Detection
- 2 Segment-based Representation
- 3 Sum-Max Video Pooling
- 4 Sum-Max Video Pooling with Better Video Segmentation
- 5 Next Study

Multimedia Event Detection

Motivation



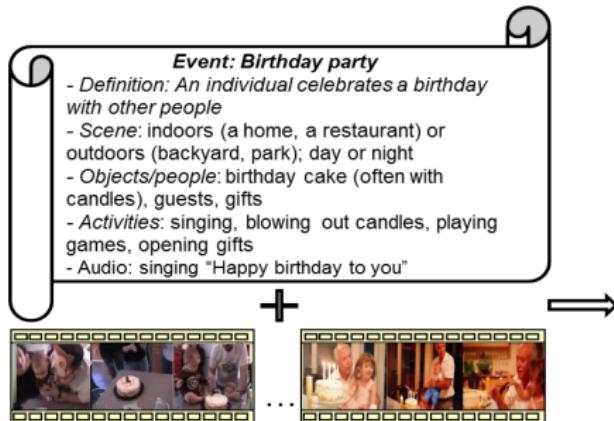
- Massive number of videos are produced every day.
 - ▶ YouTube: 72 hours uploaded per minute, with 3 billion viewers a day.
- Video need to be indexed, searched based on its content.
- Many applications:
 - ▶ User demands: tutorial videos such as "**how to make a cake**", "**how to repair an appliance**".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



unrelated segment

(a)

related segment



unrelated segment

related segment

(b)

unrelated segment

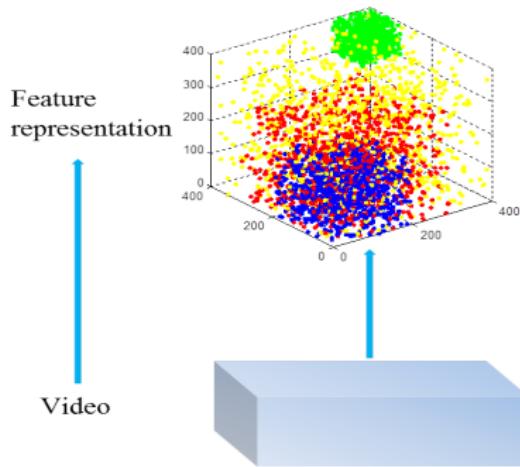
(a) Example video for "making a sandwich" event: the related segment appears after a self-cam segment (unrelated); (b) example video for "grooming an animal" event: related segment is sandwiched between two unrelated segments. This kind of video is popular in realistic dataset.

Table of Contents

- 1 Multimedia Event Detection
- 2 Segment-based Representation
- 3 Sum-Max Video Pooling
- 4 Sum-Max Video Pooling with Better Video Segmentation
- 5 Next Study

Video-based Approach

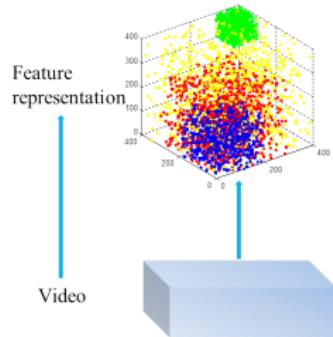
- Features are computed over the whole video
- One representation for each video



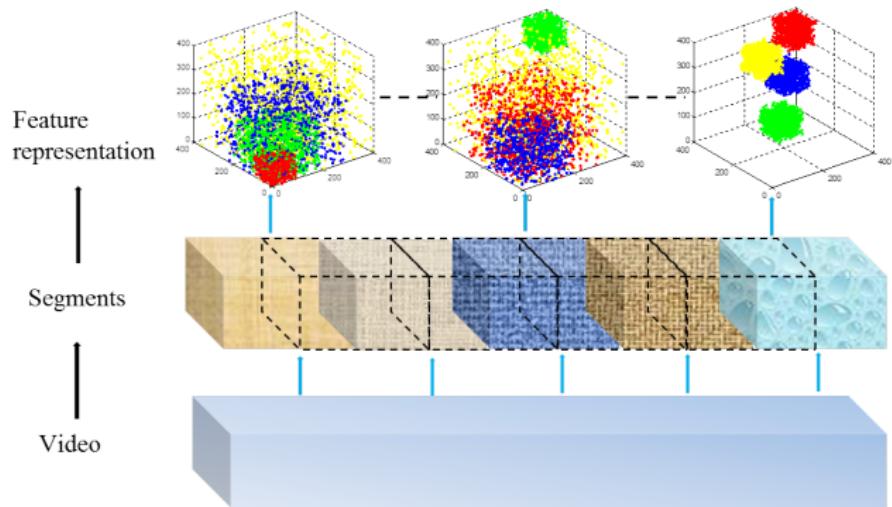
- Used by best MED'10 system (Columbia University)
- Used by best MED'11 system (BBN VISER)

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

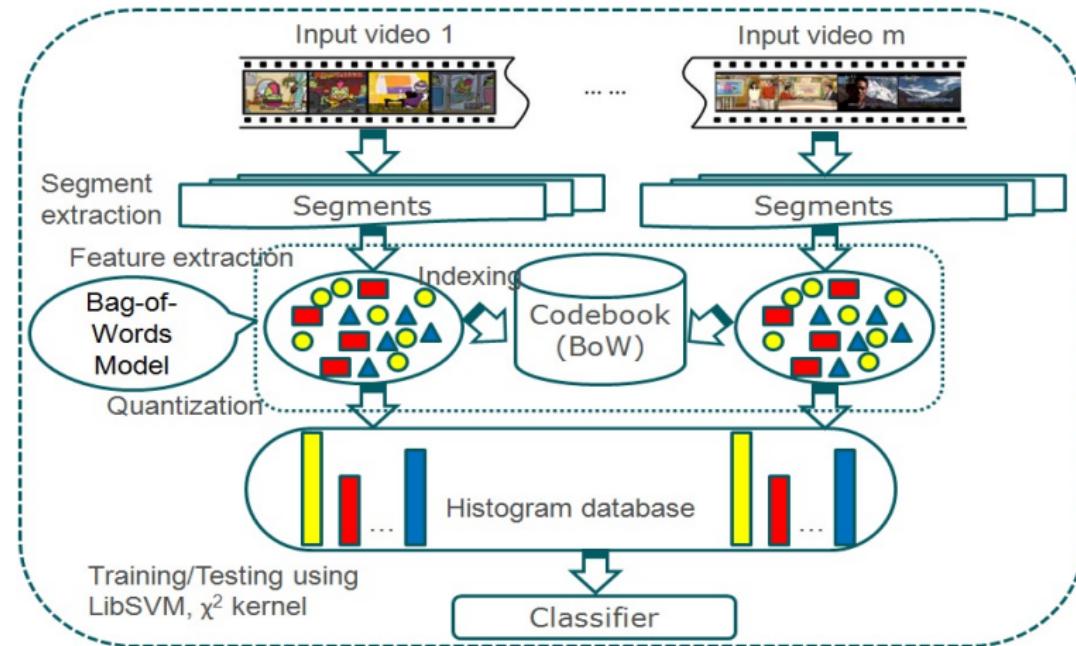
Our Segment-based Approach

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]

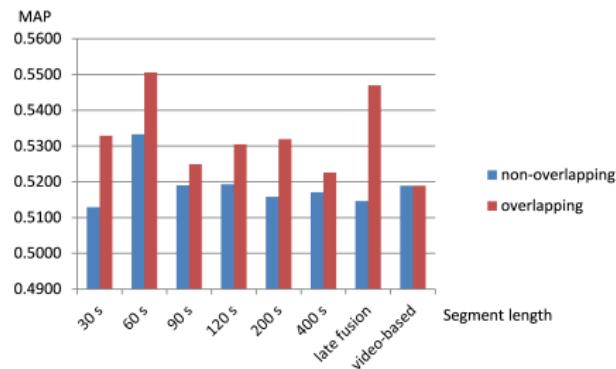
Guimaraes, S.J.F., Couprie, M., Araujo, A.d.A., Leite, N.J: Video segmentation based on 2d image analysis. Pattern Recognition Letters, 2003, 24(7), 947-957.

Evaluation Framework

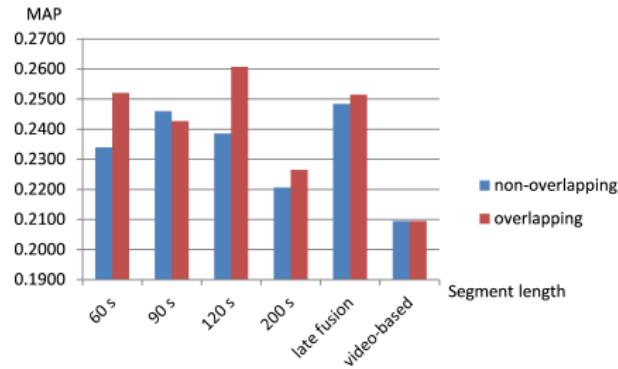


Evaluation framework for our baseline MED system

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 overlapping	SBD segments	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

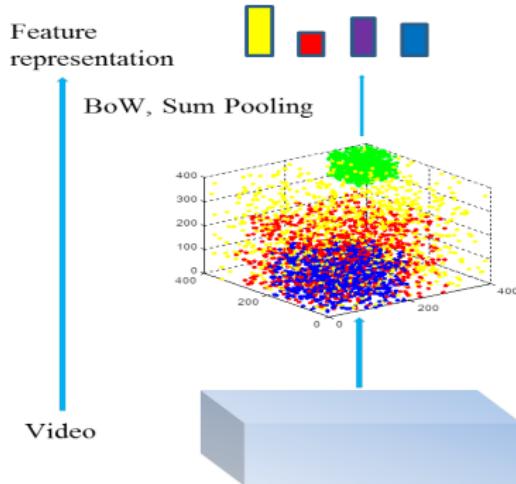
Segment-based approach outperforms the traditional video-based approach.

Table of Contents

- 1 Multimedia Event Detection
- 2 Segment-based Representation
- 3 Sum-Max Video Pooling
- 4 Sum-Max Video Pooling with Better Video Segmentation
- 5 Next Study

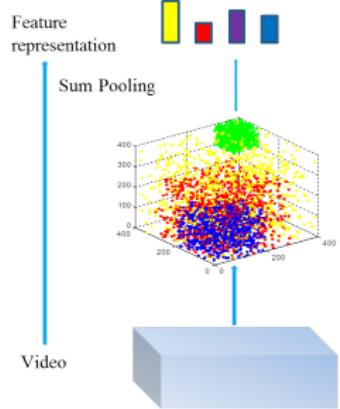
Video-based Approach

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

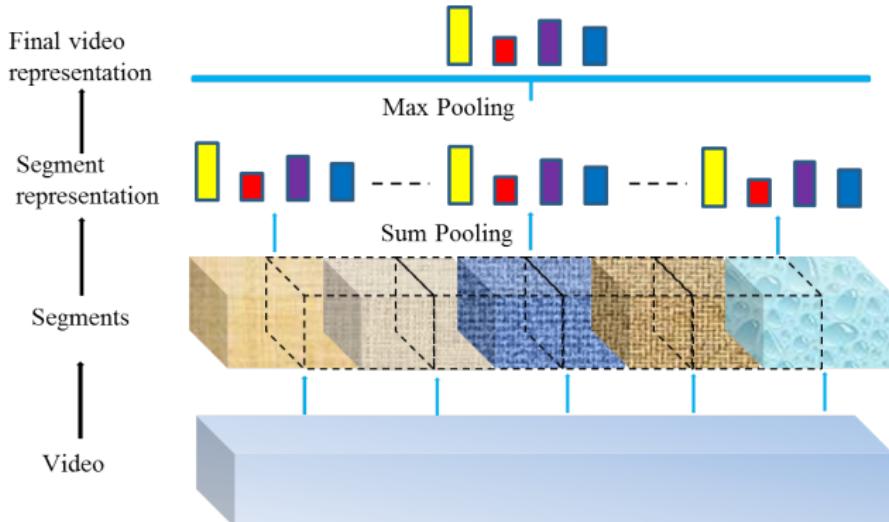


- The proposed segment-based approach: increasing the number of segment representation is not scalable.
- **Specific problem:** How to generate one representation per each video from its segment-level representations?

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 $n = 1, \dots, N$ and D is the feature dimension
- K visual words $m_k \in R^D$, where $k = 1, \dots, K$
- $M = \{m_k\}$ is the set of visual words
- Coding step: $\phi_n = [\Phi_{1n}, \dots, \Phi_{Kn}]$
- 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:

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

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

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?

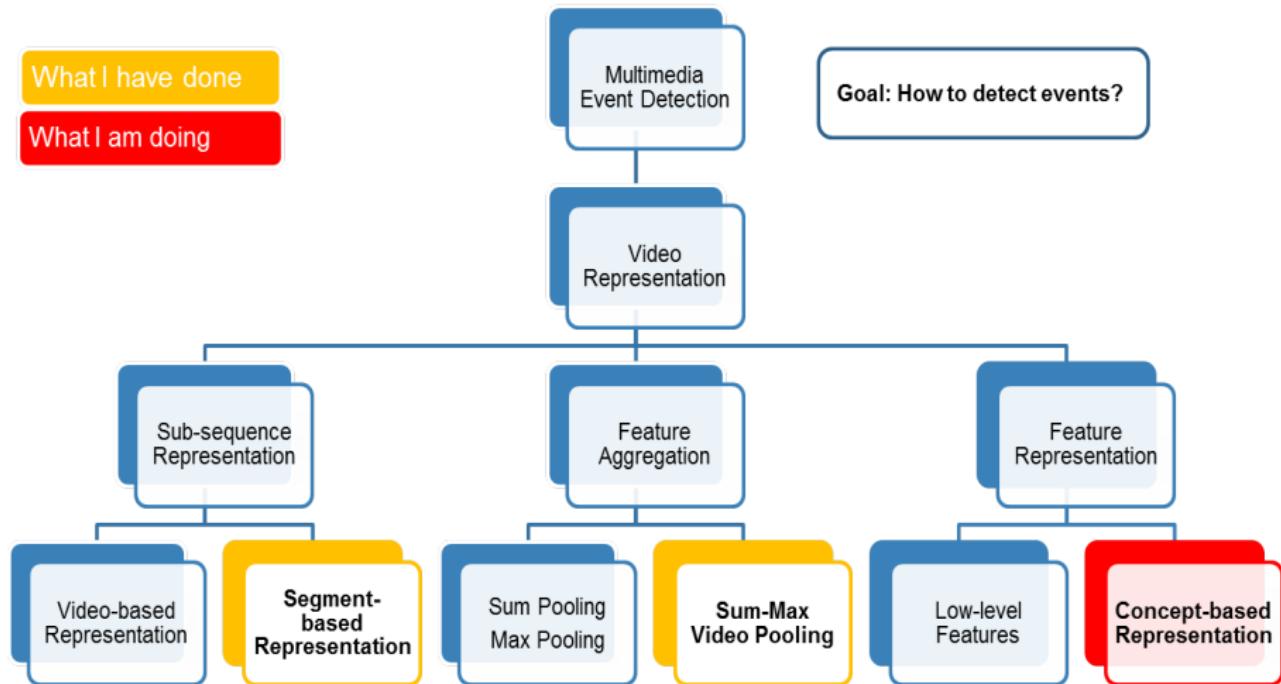


Table of Contents

- 1 Multimedia Event Detection
- 2 Segment-based Representation
- 3 Sum-Max Video Pooling
- 4 Sum-Max Video Pooling with Better Video Segmentation
- 5 Next Study

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.

A Better Way to Segment the Video

- Shot Boundary Detection.



Where am I now?

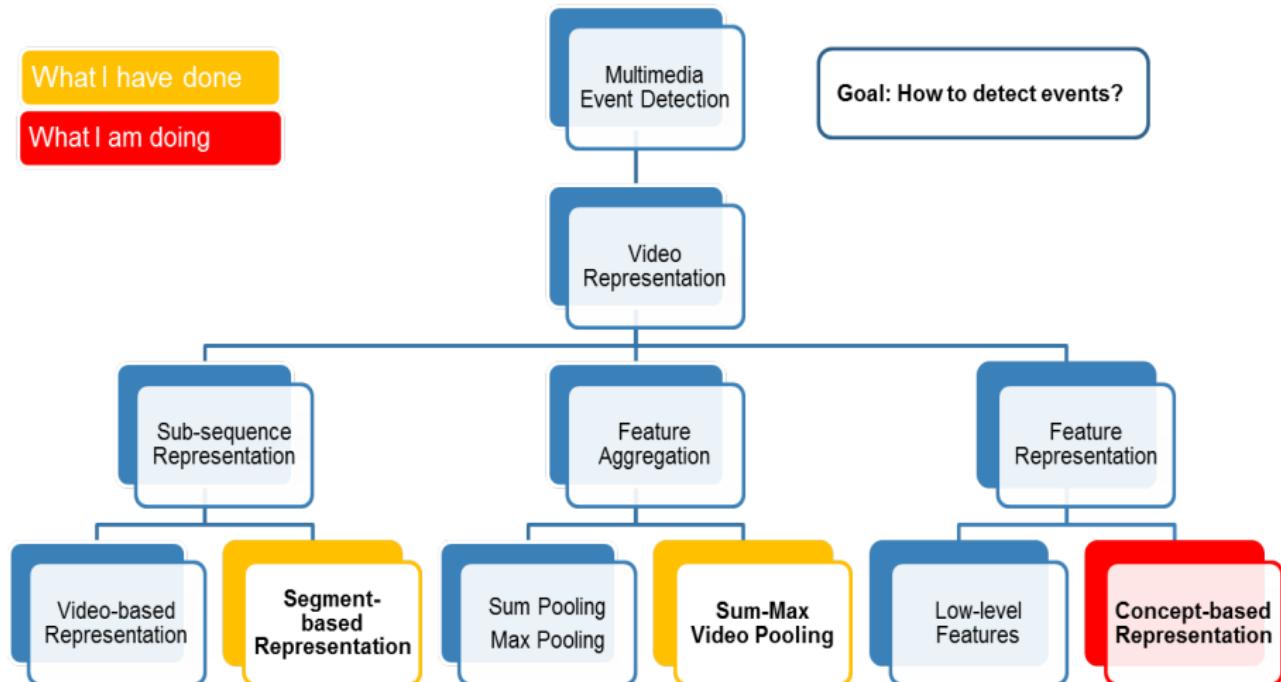


Table of Contents

- 1 Multimedia Event Detection
- 2 Segment-based Representation
- 3 Sum-Max Video Pooling
- 4 Sum-Max Video Pooling with Better Video Segmentation
- 5 Next Study

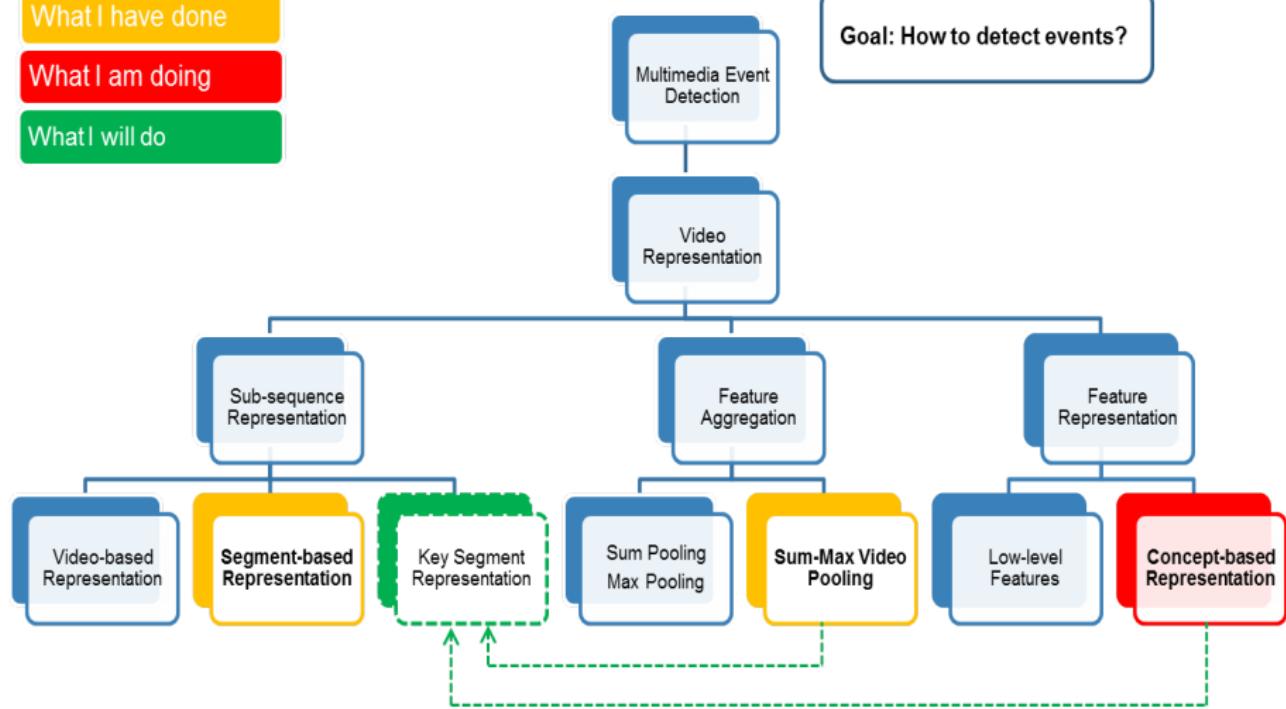
Next Study

What I have done

What I am doing

What I will do

Goal: How to detect events?



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 → making it prone to noise.

For our segment-based approach: features are aggregated over the uniform sampled segments → might not contain key segments

Research Problem: How about automatically finding the key segments that contain positive evidence for a specific event?

First Approach

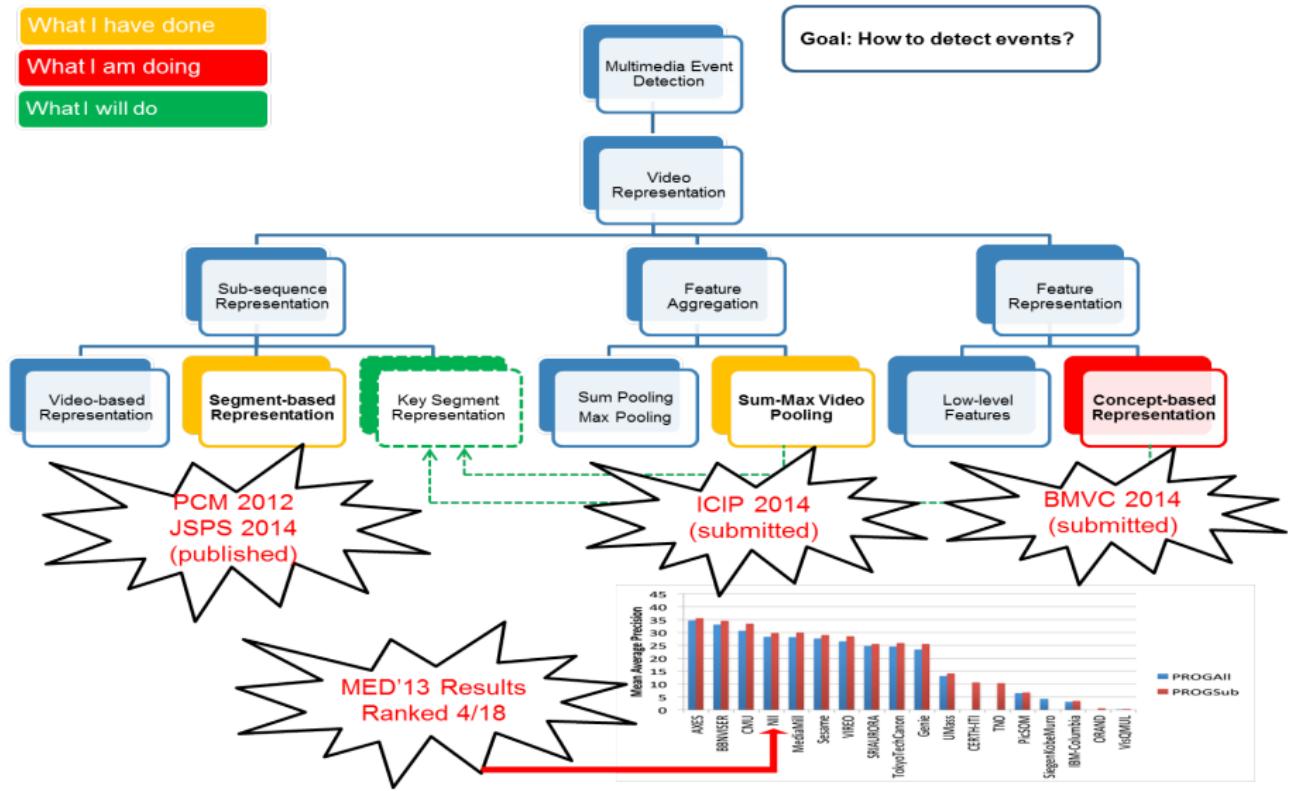
- We first analyze videos of specific event kit and manually identify key video segments to well represent that event type ()
 - ▶ E.g., "no dance" segment followed by "dance by group of people" segment for "flash mob gathering" event
- Train classifiers for each types of video segments (using current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



Thank you for your attention!