Event Detection from Video Using Segment-Based Approach

(Final Defense)

Sang Phan

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

July 22nd, 2015

Response to Committee's Requests

1. Modify the thesis title.
   → Done.
   - Old title: Multimedia Event Detection Using Segment-Based Approach.
   - New title: Event Detection from Video Using Segment-Based Approach.
2. Modify chapter titles.
   → Done.
   - Chapter 3: Event Detection Using Segment-based **Feature Representation**.
   - Chapter 4: Event Detection Using Sum-max **Feature Aggregation**.
   - Chapter 5: Event Detection Using Event-Driven Multiple Instance **Learning**.
3. Modify the story to cover three contributions.
   → Done (modified in the abstract).
4. Which challenges are solved by which methods?
   → Done (added in Chapter 6).
5. Try dynamic pooling (optional).
   → Not yet.

Table of Contents

Table of Contents

Motivations

- Massive number of videos are produced every day.
  - YouTube: 300 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**".

Motivations (cont'd)

A real world application: **detect shoplifting** ("manbiki").

- Jan, 2015: National manhunt in Japan for a YouTuber who allegedly stole...snacks [Mainichi News].

The footage shows that he's stolen many things.

- How can the police know?
  - Because he has been uploading clips of his alleged thefts.
- Can it be **automatically** detected by security cameras?

Complex Event

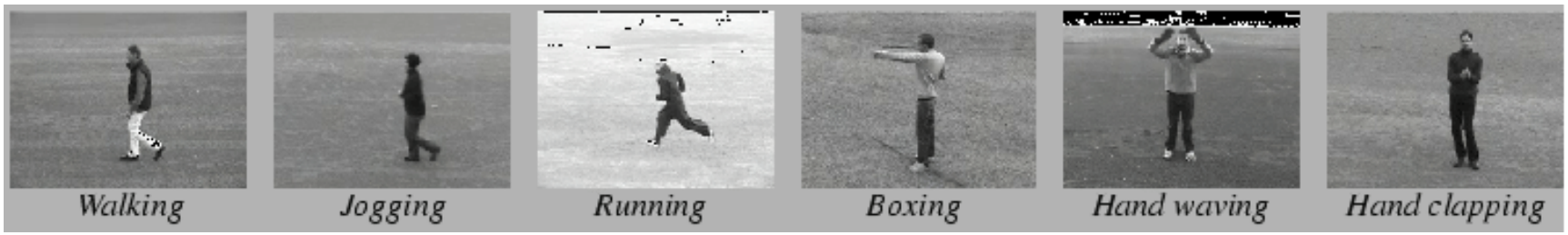
Definition

- is a complex activity occurring at a specific place and time;
- involves people interacting with other people and/or objects;
- consists of a number of human actions, processes and activities.

Pick up an item | Keep it in a hidden place | Get out successfuly

Sequence of actions in the shoplifting event.

Compared to **single action** detection [KTH dataset]

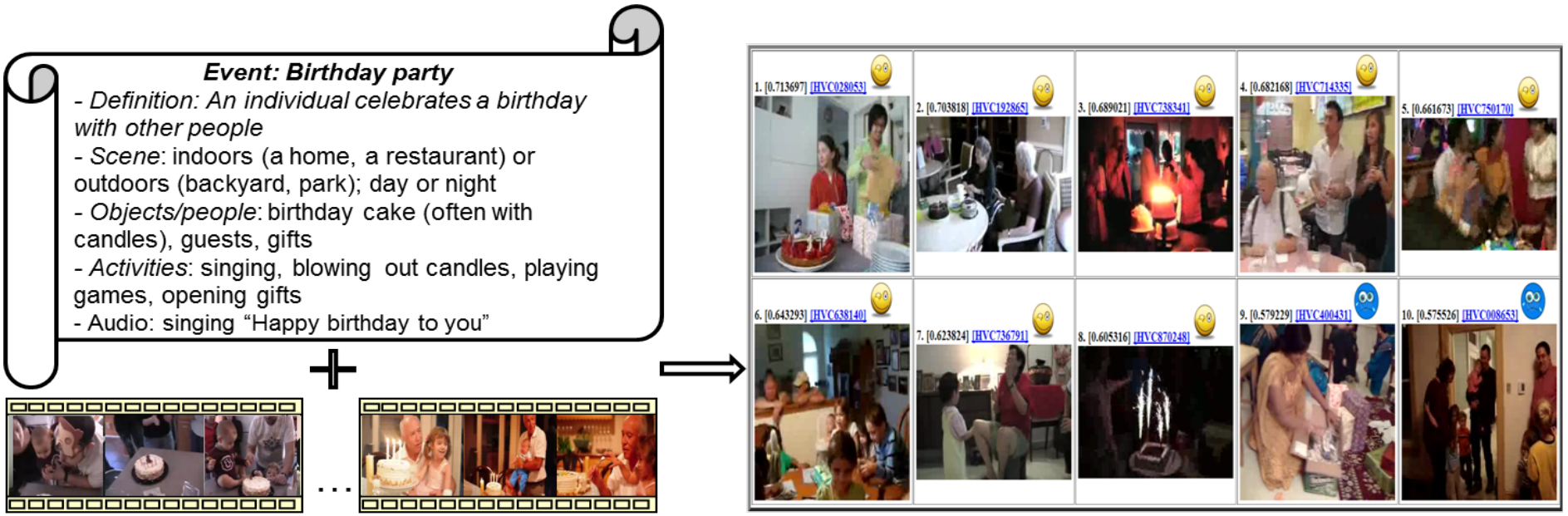


Event Detection from Video

In 2010 TRECVID proposed Multimedia Event Detection (MED) task.

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.



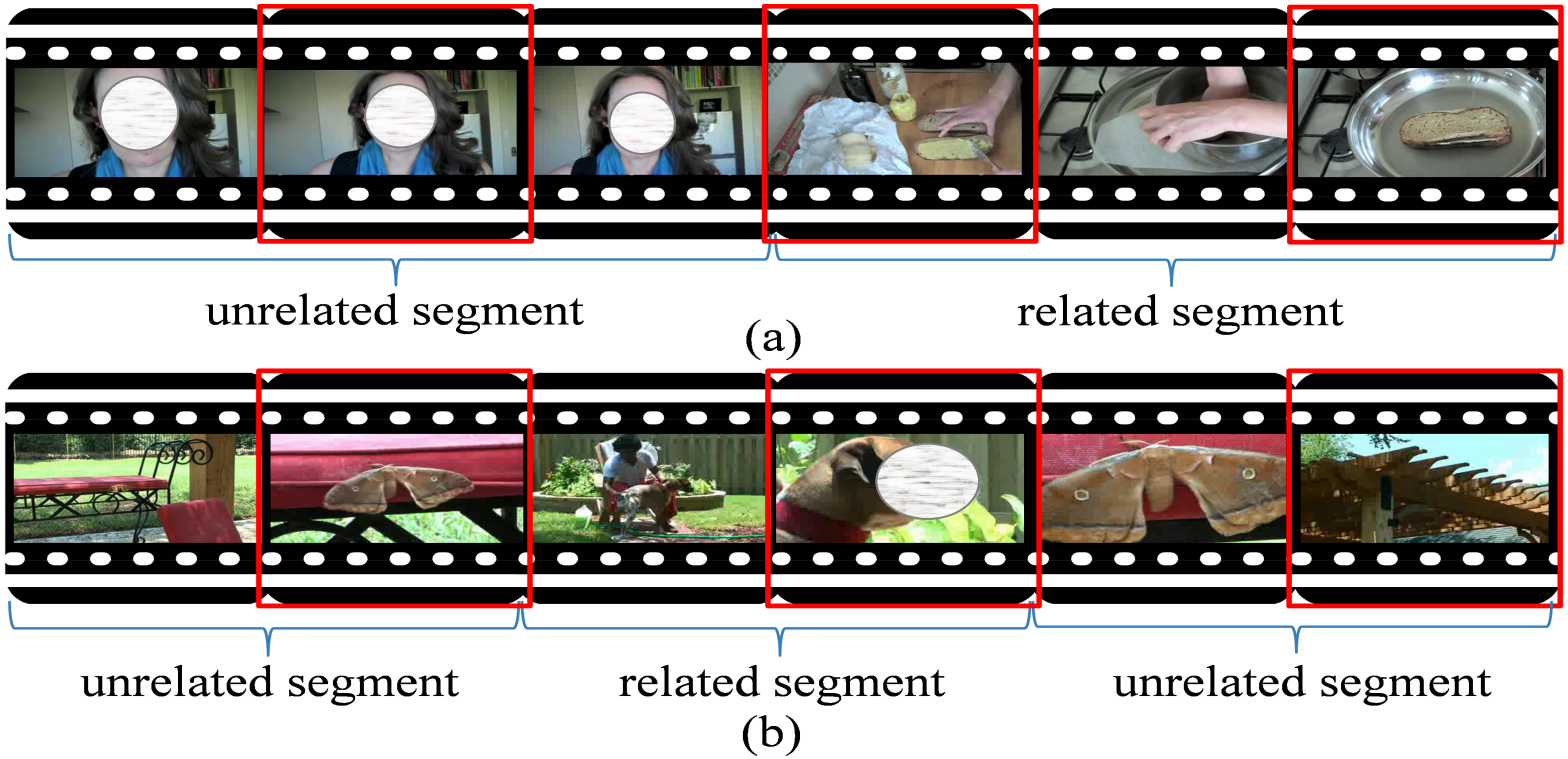
Challenges of Event Detection from Video

- **Large content variation**: the diversity of complex event is very high.

The large variation of *birthday cake* in the "birthday party" event.

Challenges of Event Detection from Video (cont'd)

- **Uncontrolled capturing conditions**: different time, location, clutter in the environment, camera motion $\rightarrow$ can contain irrelevant information to the event of interest.



(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.

Challenges of Event Detection from Video (cont'd)

- Presence of **near-miss** (related) videos.
  - closely related to the event but it lacks critical evidences to be a positive event instance.

Example of *near-miss* videos for "Changing a vehicle tire" event (2nd row).

Challenges of Event Detection from Video (cont'd)

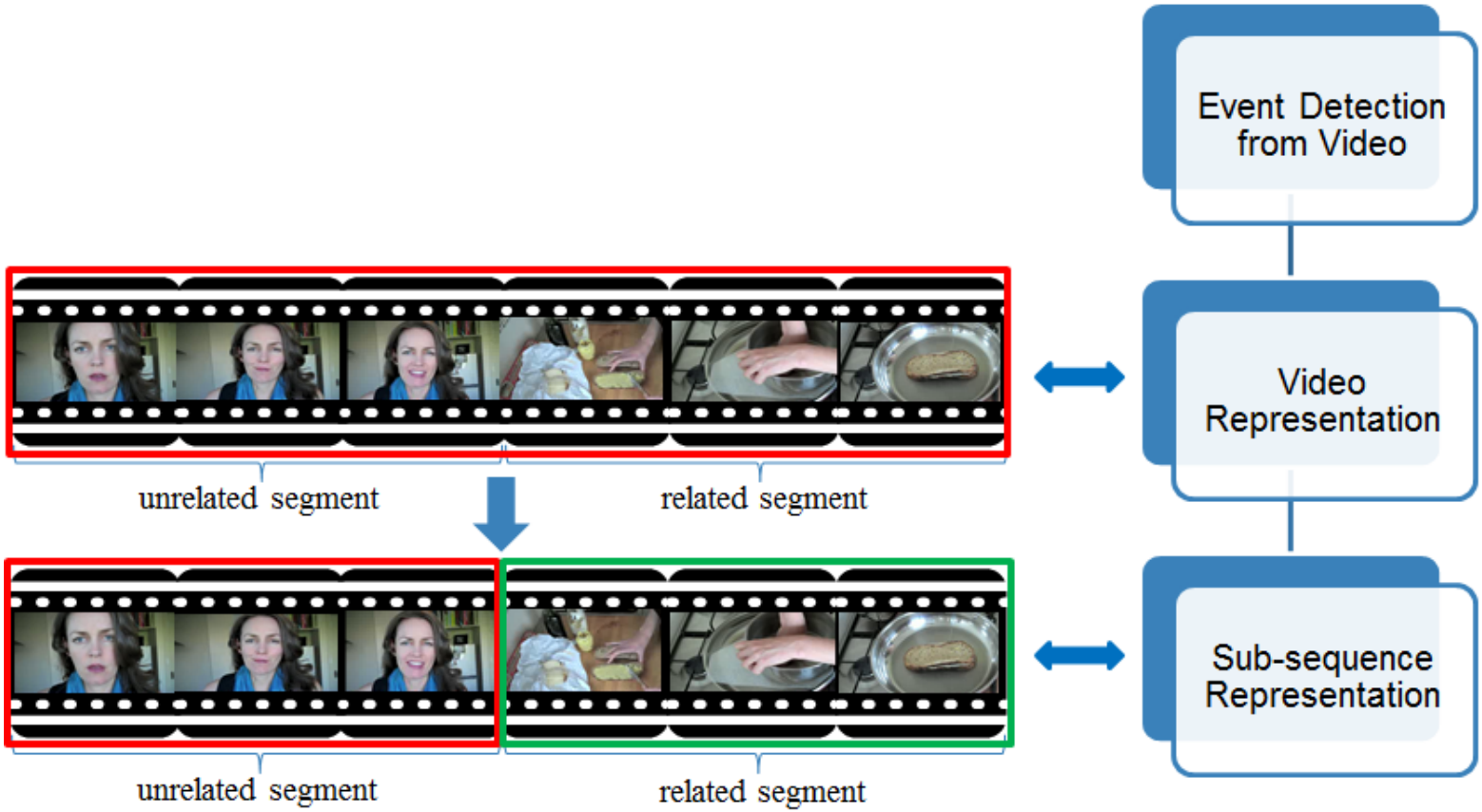
- Large scale video archives

Number of videos in the TRECVID MED collection up to 2014.

Set		Number of video clips	Video duration (hours)
Development Data	RESEARCH	10,000	314
	10 Event Kits	1,400	74
	Transcription	1,500	45
Event Training Data	Event Background	5,000	146
	40 Event Kits	6,000	270
Test Data	MEDTest	27,000	849
	KindredTest	14,500	687
Evaluation Data	MED14Eval-Full	198,000	7,580
	MED14Eval-Sub	33,000	1,244
Total		**244,000**	**9,911**
THUMOS14 (largest action dataset)		**13,000**	**254**

Target Challenge & Research Direction

- Target challenge: **Uncontrolled capturing conditions**
  - contains irrelevant information to the event of interest.
  - differentiates real video from studio/controlled-capturing video.
- Research direction: Decompose the video into sub-sequences and study event detection methods from these sub-sequences.



Contributions

1. Segment-based **Representation** (SB)
   - Investigate different strategies to decompose a video into segments.
   - Study the optimal segment length.
2. Sum-Max Video **Aggregation** (SM)
   - An efficient method to aggregate local features into video feature representation.
3. Event-driven Multiple Instance **Learning** (EDMIL)
   - A method to leverage the event description to learn key evidences for complex event detection.

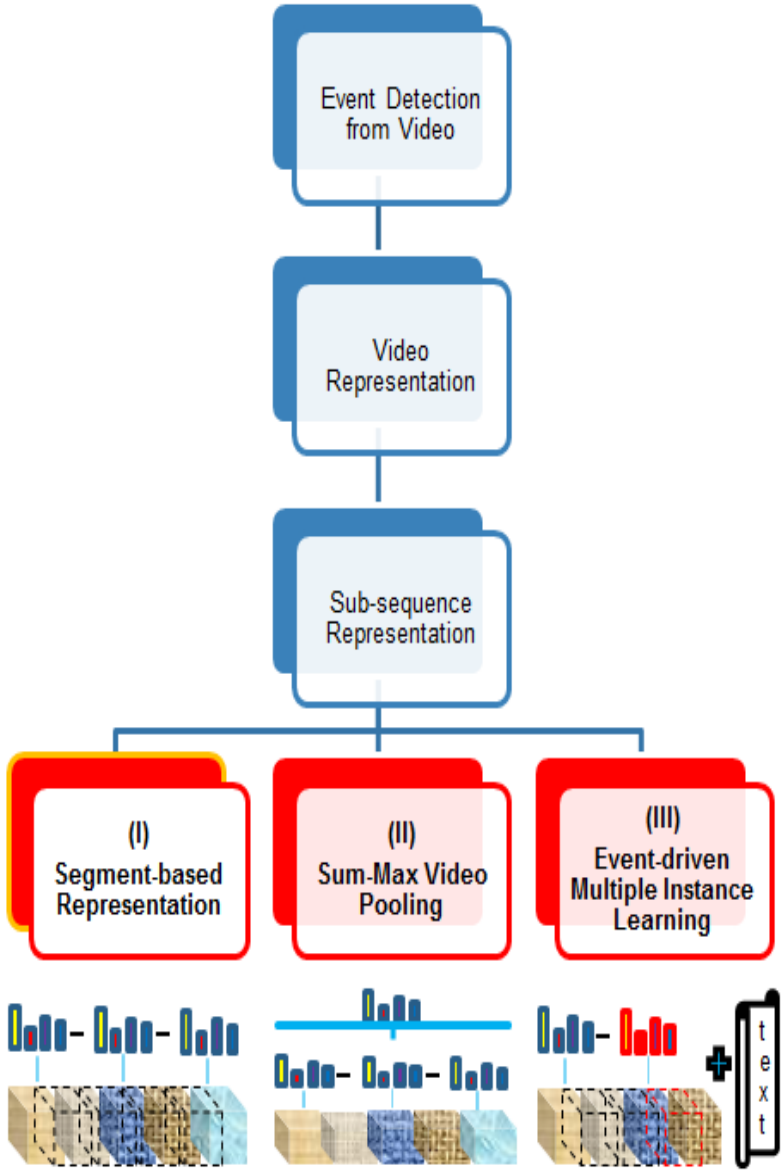
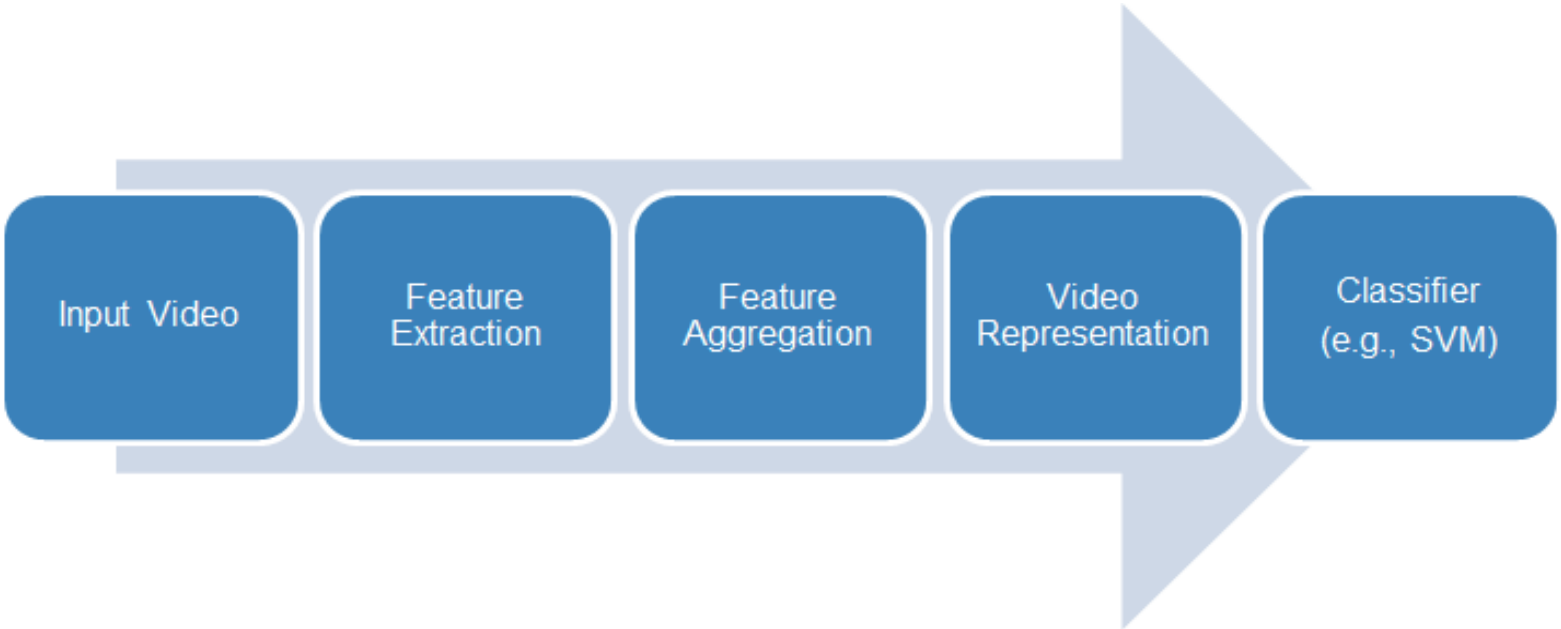


Table of Contents

How to Detect Event in Video?

- State-of-the-art systems [Jiang-TRECVID2010], [Natarajan-TRECVID2011] (**best systems** in TRECVID MED 2010 and 2011)



Problem: Neutralize the contribution of each part of the whole video.

- In fact, the clues to determine an event often appear in a small segment.

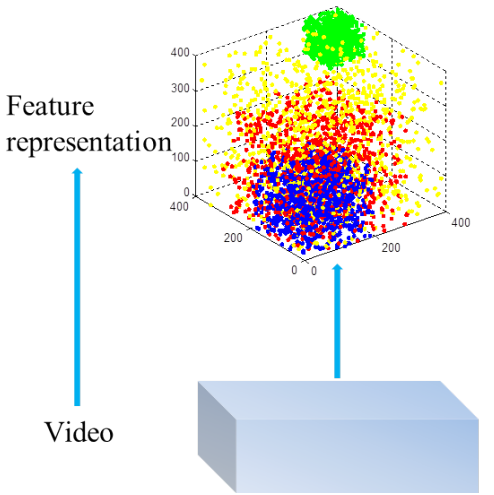
How to Detect Event in Small Segments of Video?

- Straightforward solutions: Split the videos into segments and detect event in these small segments.
  - [Niebles-ECCV2010] and [Gaidon-CVPR2011] model activities as sequences of atomic actions using semantic attributes.
  - [Tang-CVPR2012] model the key segments and segment duration as latent variables and solve using a variant of HMM.
  - [Vahdat-ICCV2013] Localize the most salient evidences using latent SVM.
  - [Lai-CVPR2014] Detect key instances in video based on a variant of Multiple Instance Learning.
- Limitations
  - Segmentation of activities into predefined atomic actions and annotation of these atomic actions are available.
  - Video sequences are precisely cropped with activities of interest.
  - Use fixed length segment.

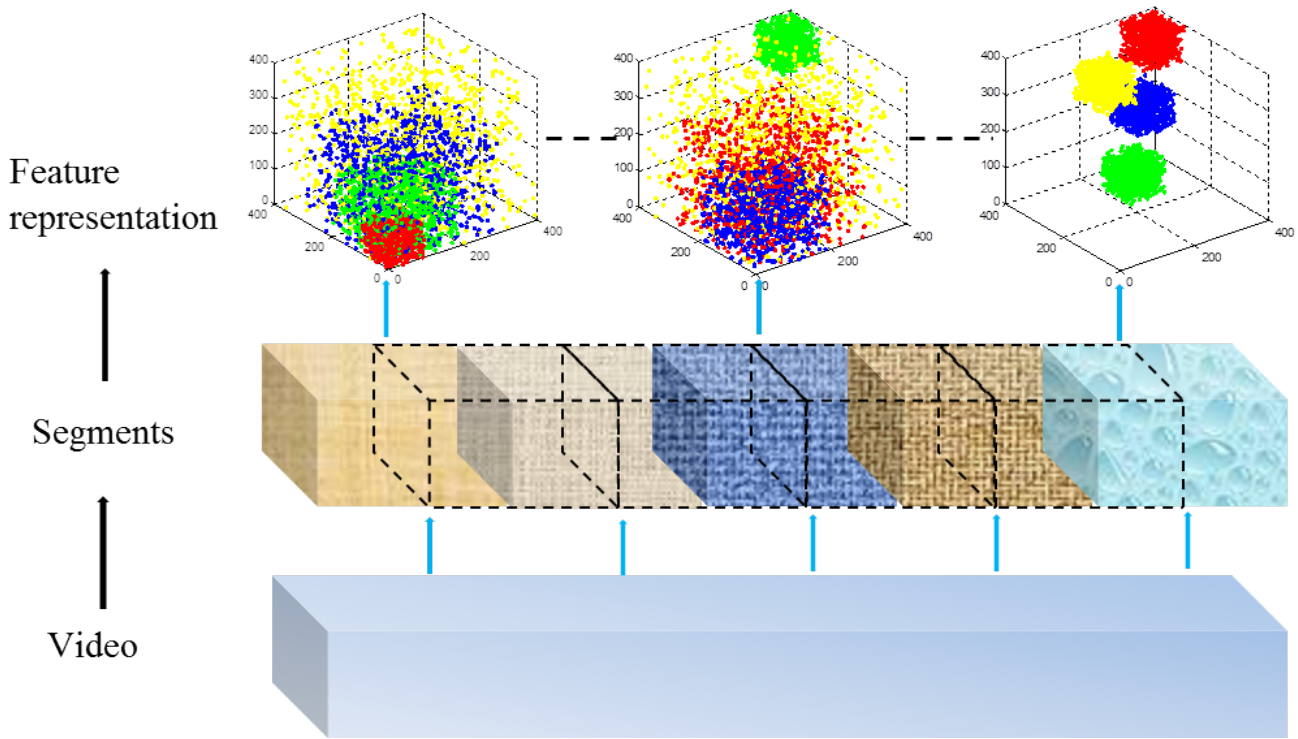
$\rightarrow$ **We do not know what is the optimal segment length when splitting the whole video!**

Our Segment-based Approach

- We investigate different strategies to split the videos in segments.
- We study the optimal segment length.



(a) The video-based approach



(b) **Our proposed segment-based approach**

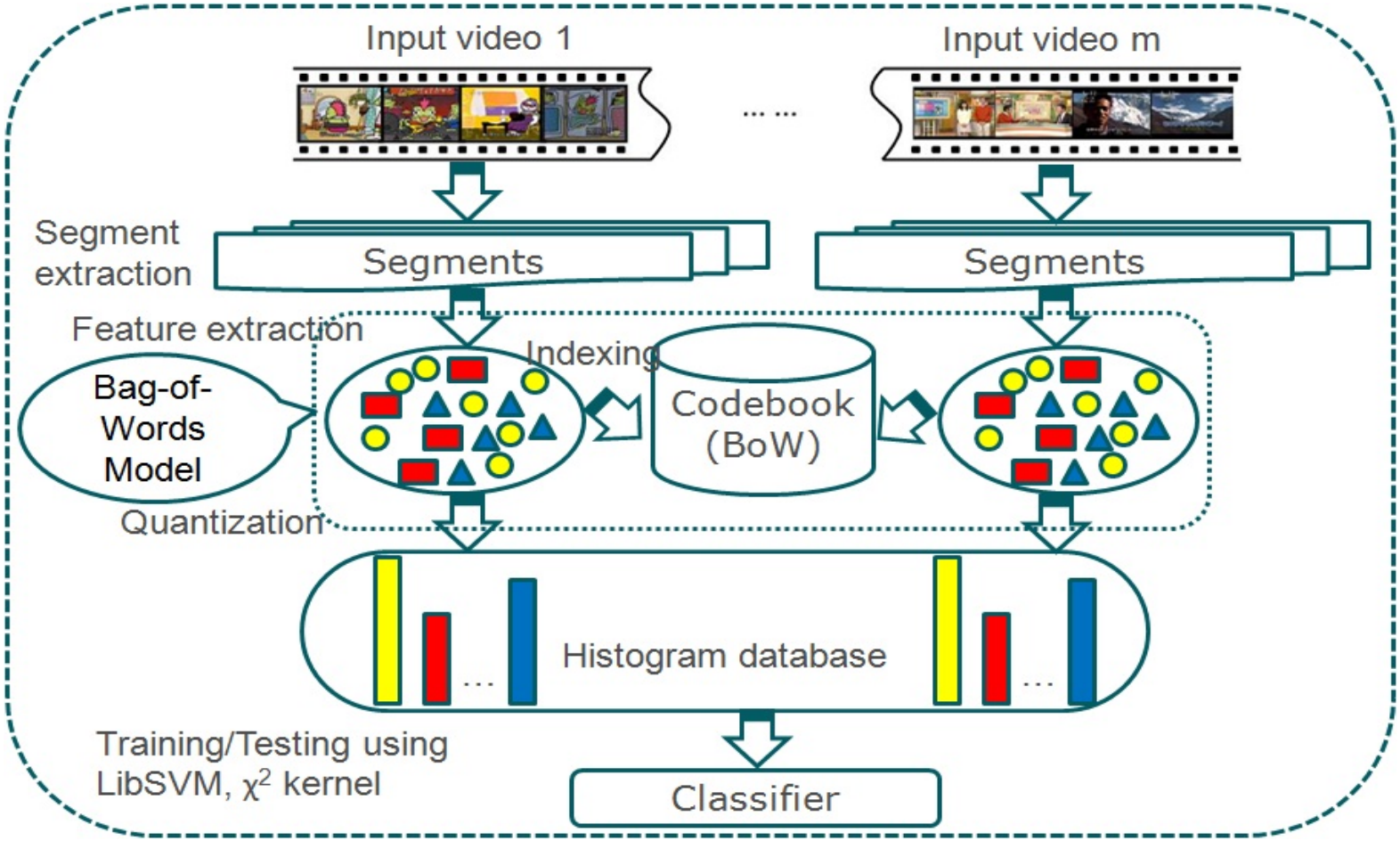
Our Segment-based Approach

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 event detection system.

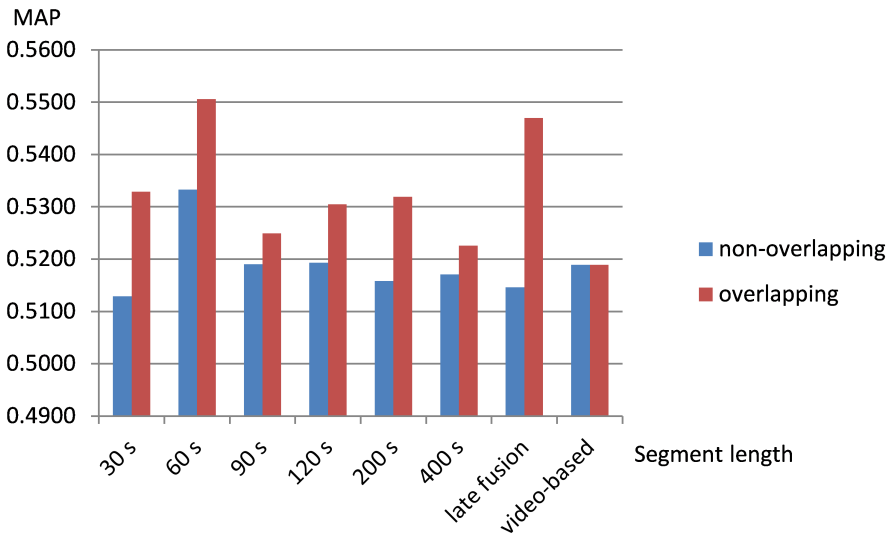
Experimental Setup

- Dataset

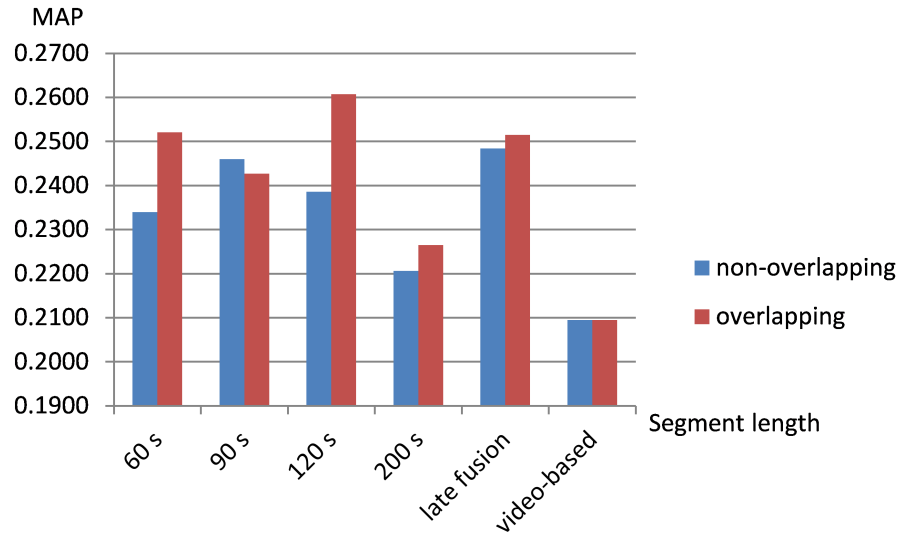
Dataset	No. Event	No. Train Videos	No. Test Videos	Total Videos	Total Hours
MED2010	3	1,744	1,724	3,468	110 hours
MED2011	10	12,590	31,822	33,153	1,400 hours
MED2012	25	3,878	1,938	5,816	250 hours

- Feature: Dense Trajectories, MBH descriptor [Wang-CVPR2011]
- Feature encoding: Bag-of-words model, 4000 codewords.
- Learning: χ^2 SVM.

Result: Non-Overlapping vs. Overlapping Sampling



(b) On the MED 2010 dataset



(b) On the MED 2011 dataset

- Segment-based approach significantly outperforms the Video BoW baseline.
- In most cases, the overlapping sampling performs the best.

Experimental Results: On the MED 2010

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 video-based approach.
- Shot boundary detection does not work.
  - It is difficult to detect shot boundary with high accuracy in real videos.

Experimental Results: On the MED 2011

Comparison of different segment-based approaches with the video-based approach.

Event	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

- Segment-based approach outperforms the video-based approach.
- **Optimal segment length**: approximate the **mean video length**.
  - Late fusion of several runs around the mean video length.

Conclusions

1. **Segment-based Representation (SB)**
   - Investigate different strategies to decompose a video into segments.
   - Study the optimal segment length.
2. Sum-Max Video Aggregation (SM)
3. Event-driven Multiple Instance Learning (EDMIL)

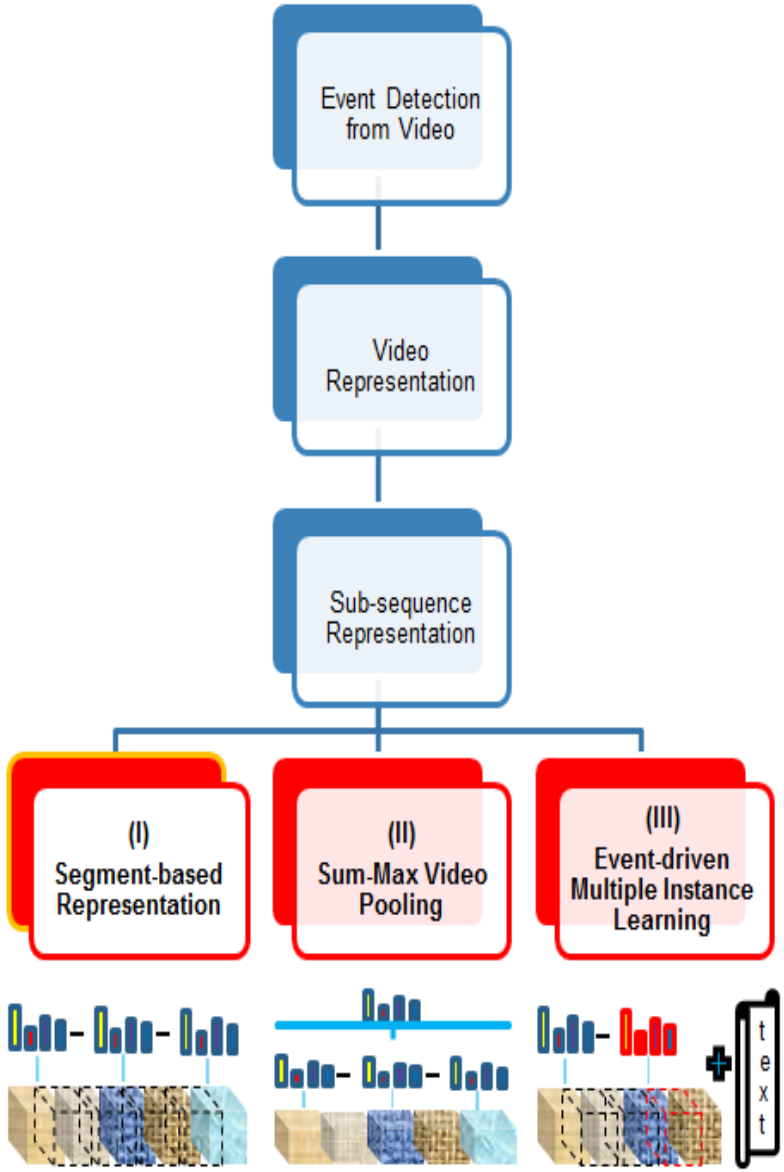
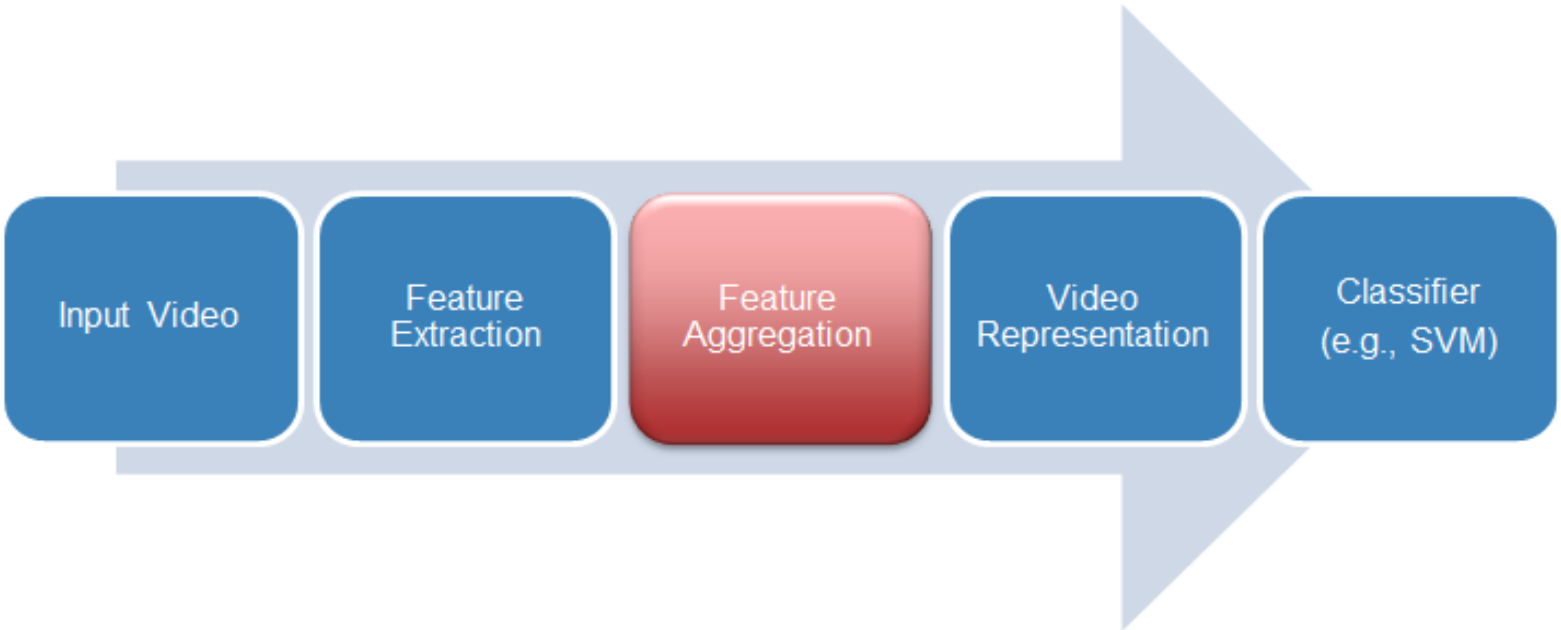


Table of Contents

How to Detect Event in Video?

- State-of-the-art systems [Jiang - TRECVID2010], [Natarajan - TRECVID2011] (**best systems** in TRECVID MED 2010 and 2011)



Problem: Neutralize the contribution of each part of the whole video.

- In fact, the clues to determine an event often appear in a small segment.

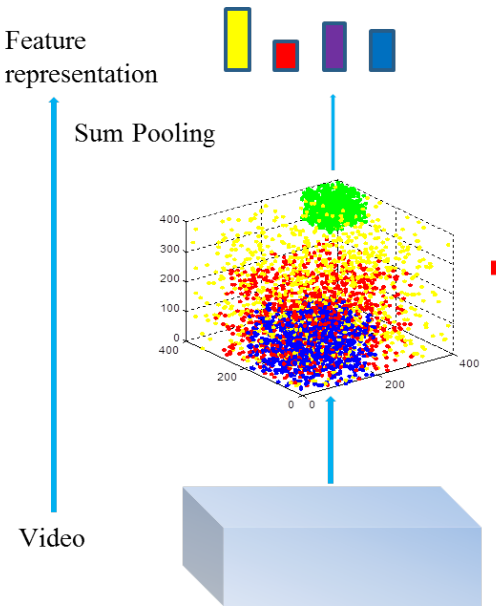
How to Aggregate Feature?

- **Sum pooling** strategy [Jiang TRECVID2010], [Natarajan TRECVID2011] (**best systems** in TRECVID MED 2010 and 2011).
  - dominant by frequently-occurring descriptors
  - rarely occurring descriptors have less influential
- **Max pooling** strategy [Wang ACCV2012] (often used with sparse coding for image classification [Yang CVPR2009])
  - only select the most discriminative information
  - likely to lost other crucial information
- **Dynamic pooling** [Li ICCV2013]: dynamically determine the pooling operator most suited for each sequence using Latent SVM.
  $\rightarrow$ very time-consuming!

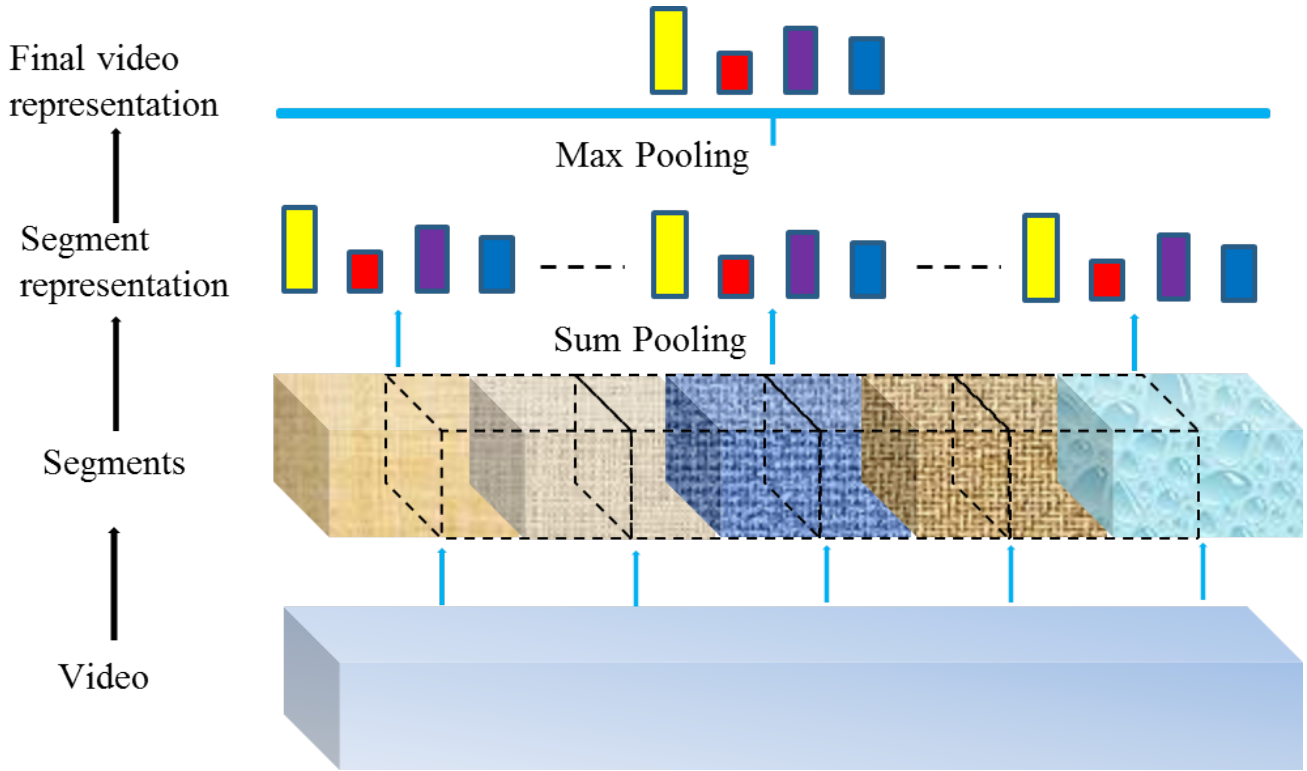
Problem: How to combine **Sum pooling** and **Max pooling** in a efficient way?

Sum-max Video Pooling

- **Sum pooling** at lower layer to accumulate sufficient features.
- **Max pooling** to retrieve the most relevant features at the high layer.
  $\rightarrow$ can discard irrelevant features in the final video representation.



(a) The video-based approach



(b) **Our proposed Sum-max Video Pooling**

Sum-max Video Pooling

- Notation:
  - N local descriptors $x_n \in R^D$, 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:

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

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

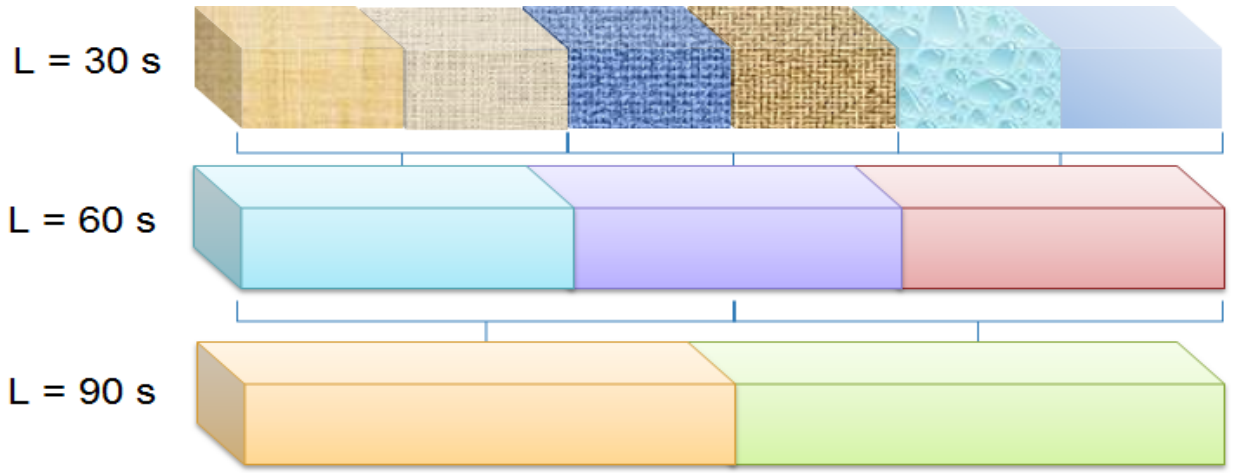
Complexity

- Same complexity with the **sum video pooling**:

$$\psi_{k_{\text{sum-max}}} = Max_{s \in S}(\sum_{n \in N_s} \Phi_{kn}) \tag{3}$$

$$\psi_{k_{\text{sum}}} = \sum_{s \in S}(\sum_{n \in N_s} \Phi_{kn}) \tag{4}$$

- Very efficient
  - Extracting features only once!
  - Aggregating features at different segment lengths efficiently.



Features from higher layers can be obtained from lower layers efficiently!

Experimental Setup

- Dataset

Dataset	No. Event	No. Train Videos	No. Test Videos	Total Videos	Total Hours
MED2010	3	1,744	1,724	3,468	110 hours
MED2011	10	1,331	31,822	33,153	1,100 hours
MED2012	25	3,878	1,938	5,816	250 hours

- Feature:
  - MED10: Dense Trajectories, MBH descriptor [Wang-CVPR2011]
  - MED11: Improved Dense Trajectories, MBH descriptor [Wang-ICCV2013]
- Feature encoding: Bag-of-words model, 4000 codewords.
- Learning: χ^2 SVM.

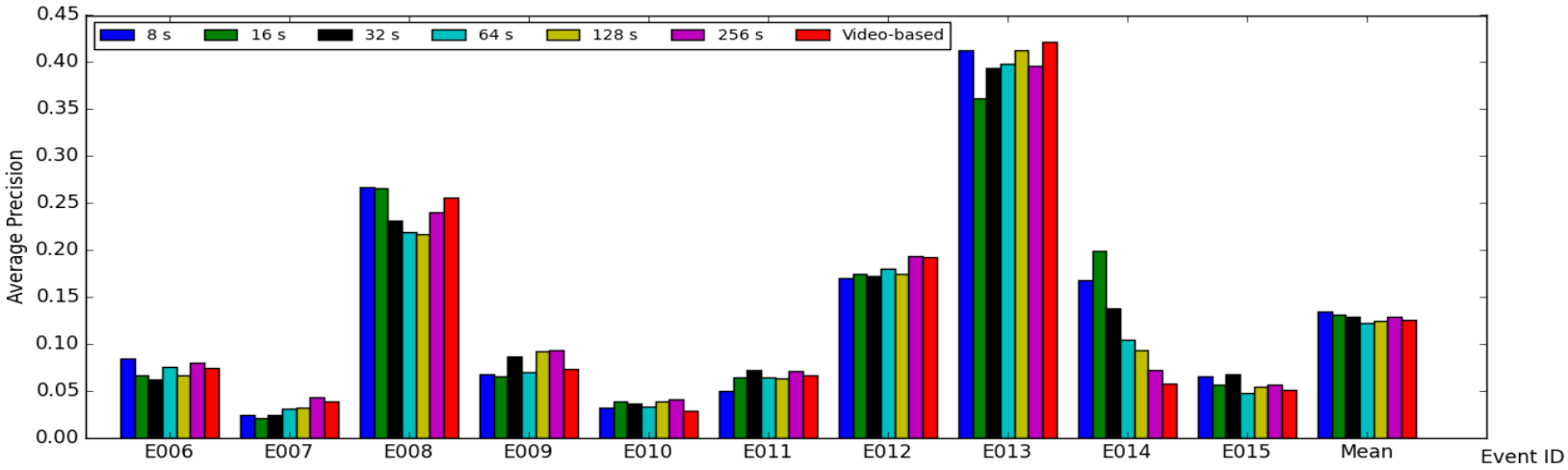
Experimental Results: On MED 2010

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.**

Experimental Results: On MED 2011



Results of Sum-max video pooling on the MED 2011 dataset.

Method/mAP	8 s	16 s	32 s	64 s	128 s	256 s	video-based
Sum-Max	**0.1339**	0.1311	0.1282	0.1220	0.1242	0.1283	0.1257
Segment-based			0.1510	0.1503	**0.1518**	0.1365	0.1257

- The best performance is at 8 s (**6**% improvement!)
- Sum-max video pooling performs significantly worse than the segment-based approach!
- However, sum-max video pooling is very **efficient**!

Conclusions

1. Segment-based Representation (SB)
2. **Sum-Max Video Aggregation (SM)**
   - An efficient method to aggregate local features into video feature representation.
3. Event-driven Multiple Instance Learning (EDMIL)

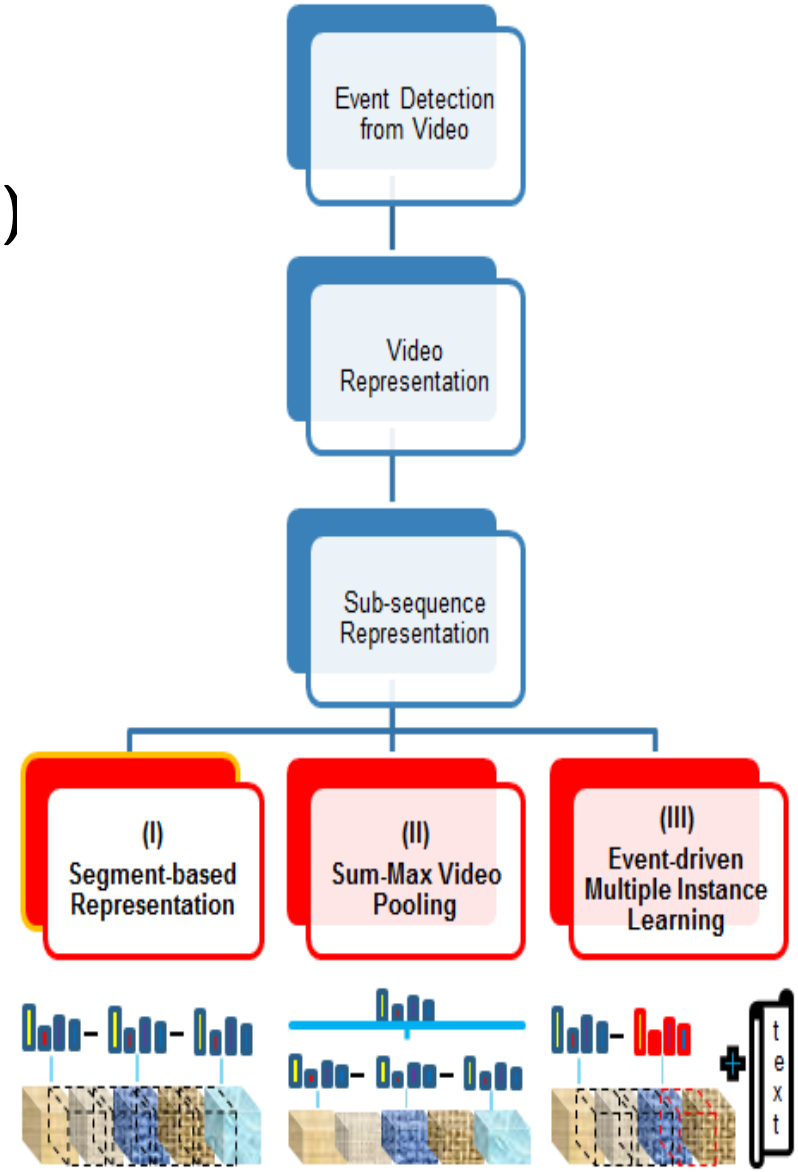


Table of Contents

Motivation - Human Way to Detect Event

- Human only needs to see some evidences.
- Leveraging on positive and negative visual cues selected by humans significantly improves the performance.

Bhattacharya, S., Yu, F. X., Chang, S. F. Minimally needed evidence for complex event recognition in unconstrained videos. In ICMR. ACM, 2014.

Motivation - Not Easy for Computer

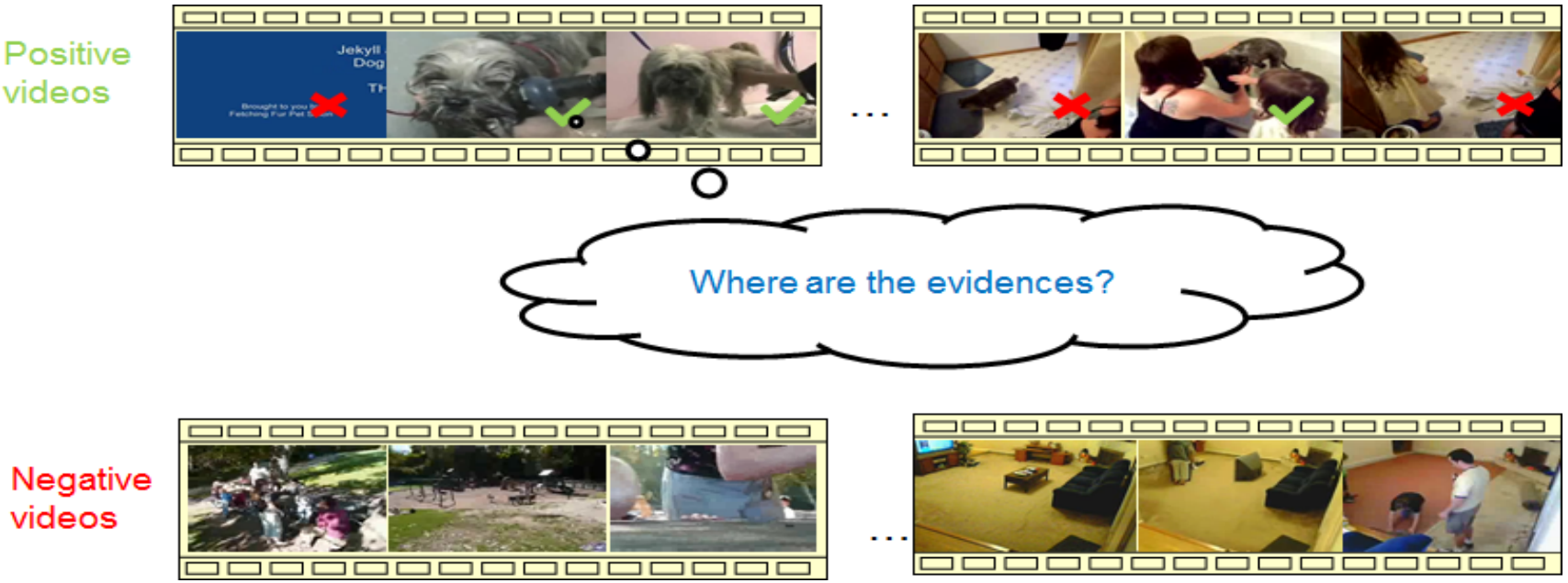
✓ Event Evidence

- How to detect which segments of the video that have the most contributions?
- How to leverage these "*key evidences*" for event detection?

→ Answering these two questions can help understand why an event is detected.

Weakly Supervised Learning Problem

Problem: Event labels are only given at video level!



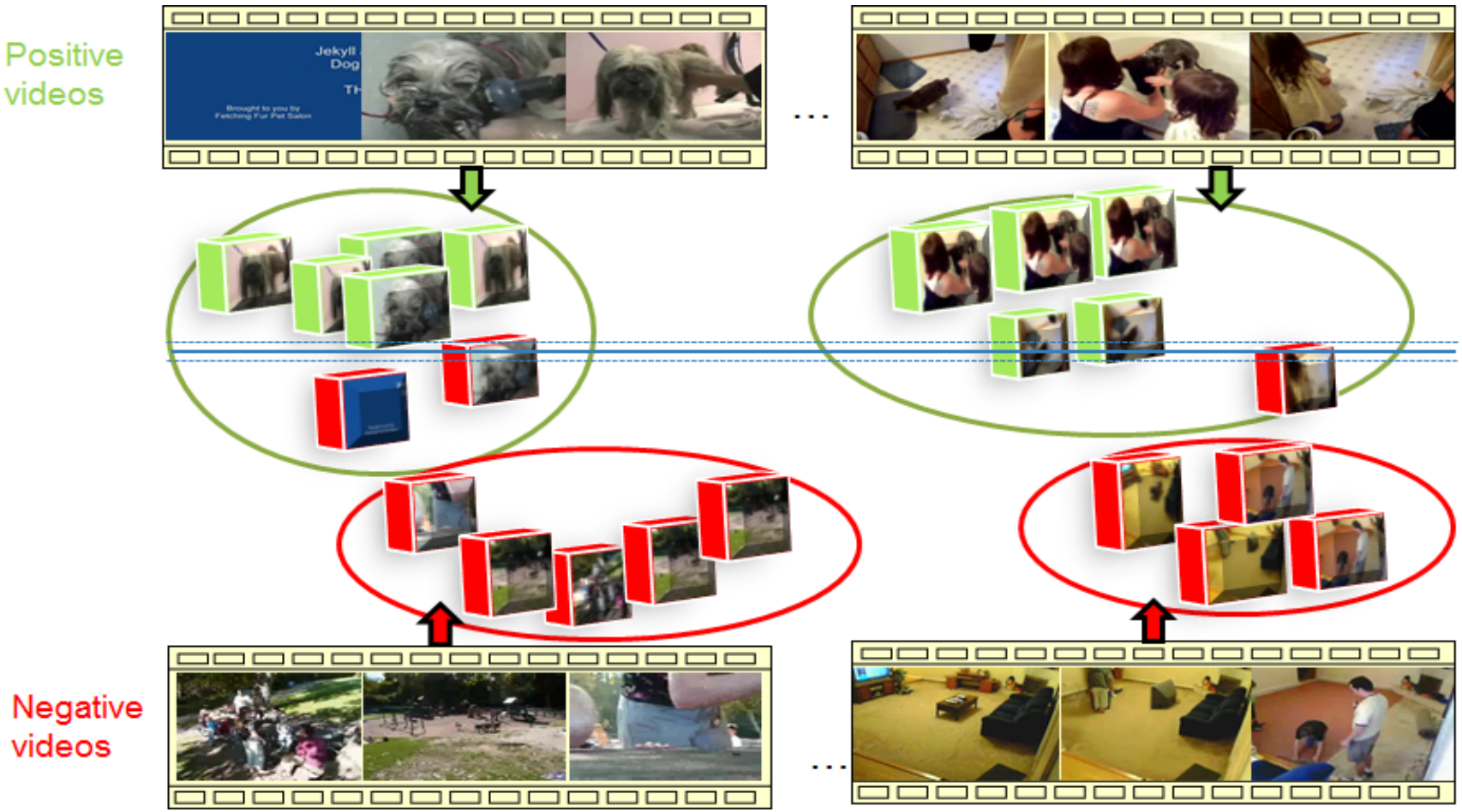
Related work

- Latent SVM [Tang-CVPR2012], [Li-ICCV2013]
- **Multiple Instance Learning** [Zhang-ICML2002], [Lai-CVPR2014]
  - Mathematically equivalent to Latent SVM!
  - MIL can be applied directly to learn labels at segment level: video $\rightarrow$ "bag", segment $\rightarrow$ "instance".

Multiple Instance Learning - miSVM

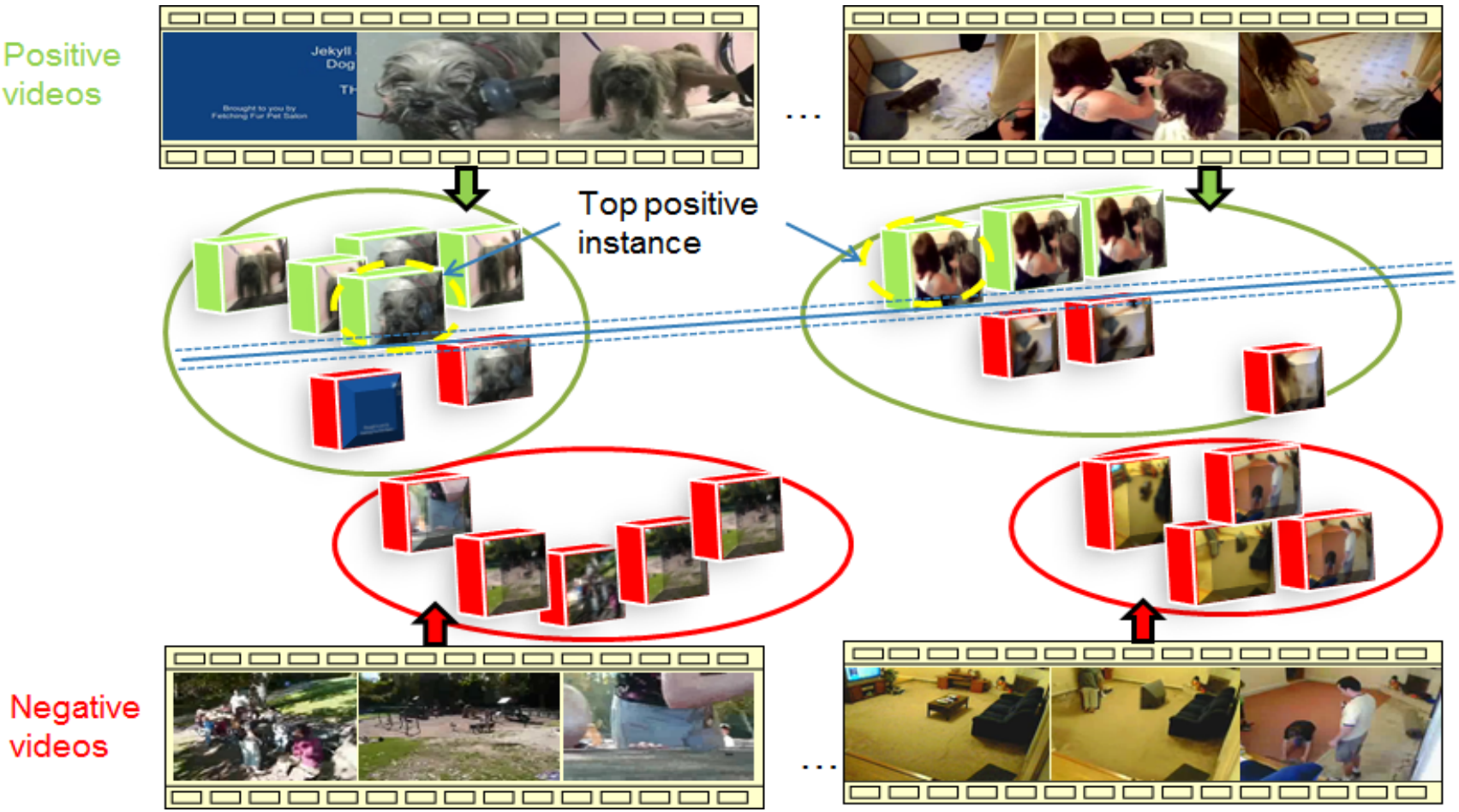
Two key assumptions of MIL:

- A positive bag has **at least** one positive instance.
- The instances in a negative bag are **all** negatives.



miSVM: instance-based classifier

Multiple Instance Learning - MISVM



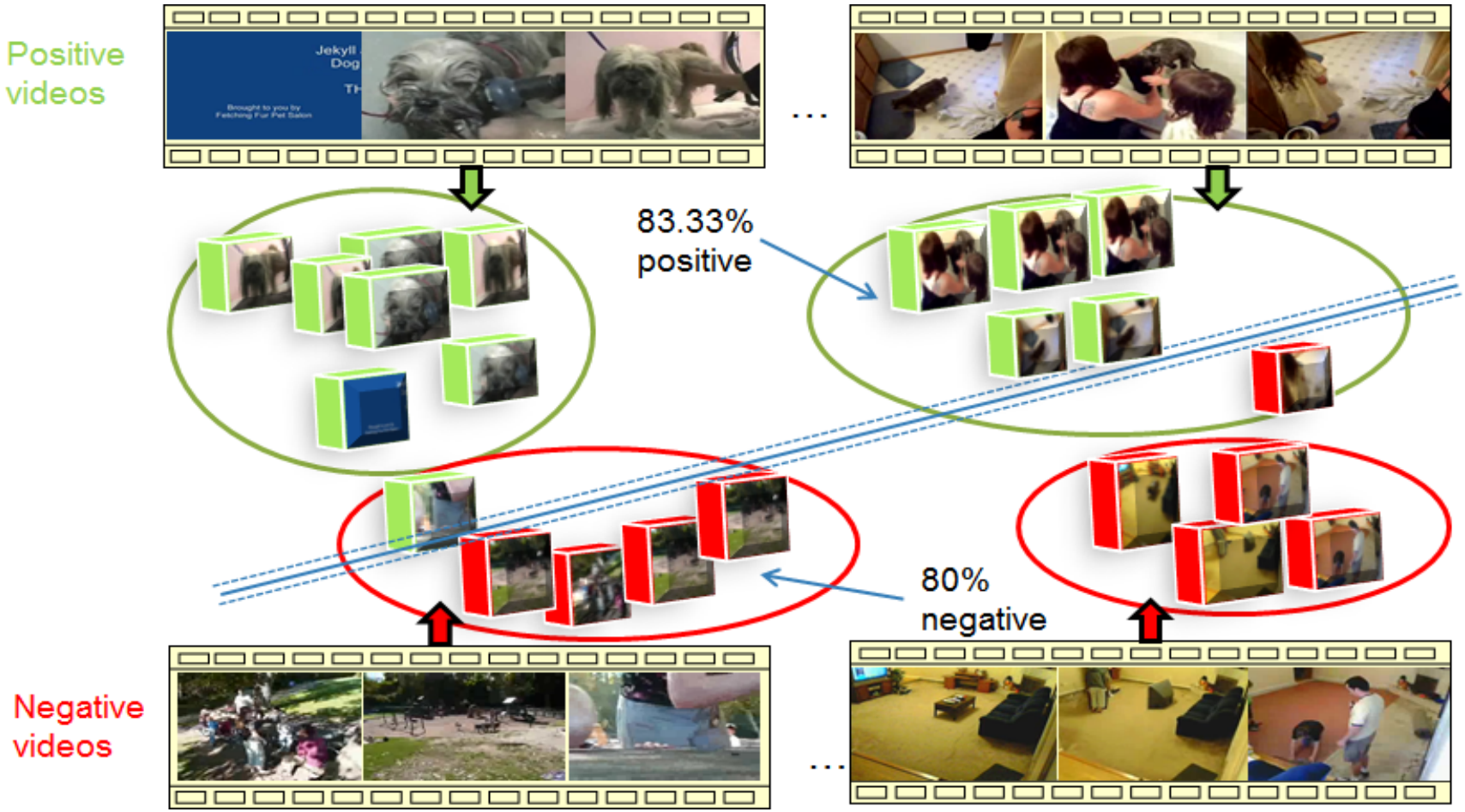
MISVM: bag-based classifier

Limitations: MIL assumptions may not be valid for videos, esp. **complex** videos.

Proportional SVM [Lai-CVPR2014]

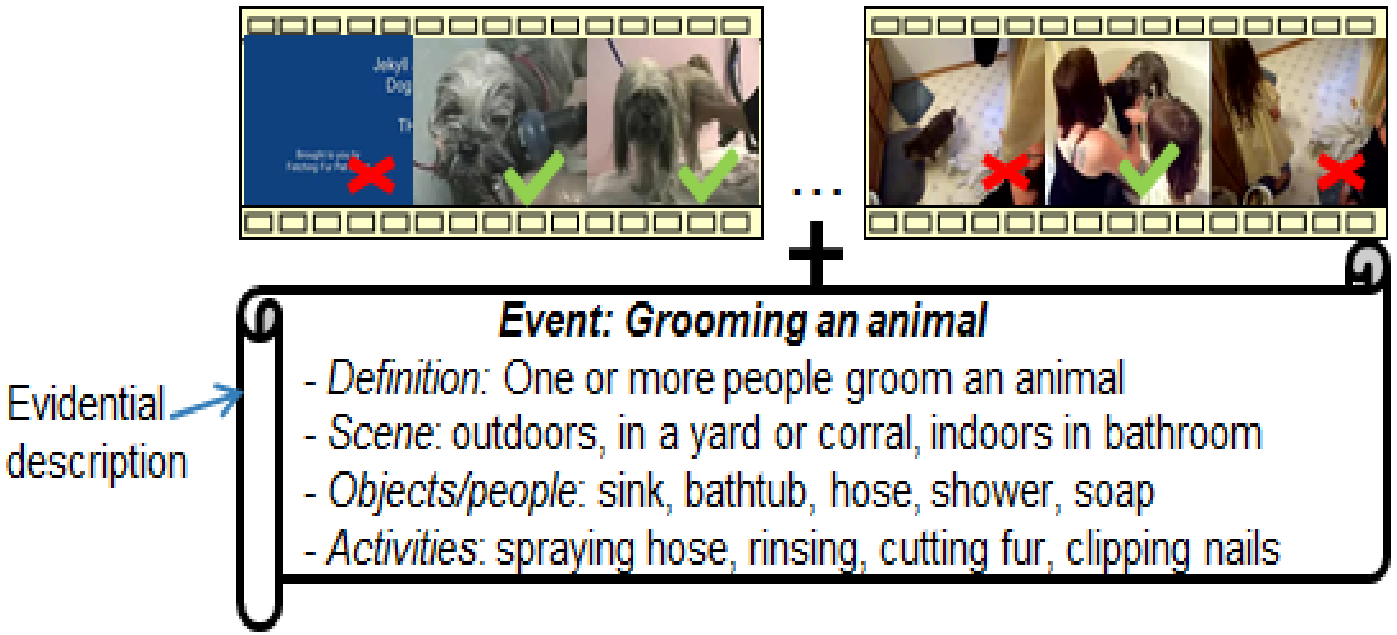
Two new assumptions of p-SVM:

- Positive videos should have **high proportion** of positive instances.
- Negative videos could have **some** positive instances.



Our Proposed Method

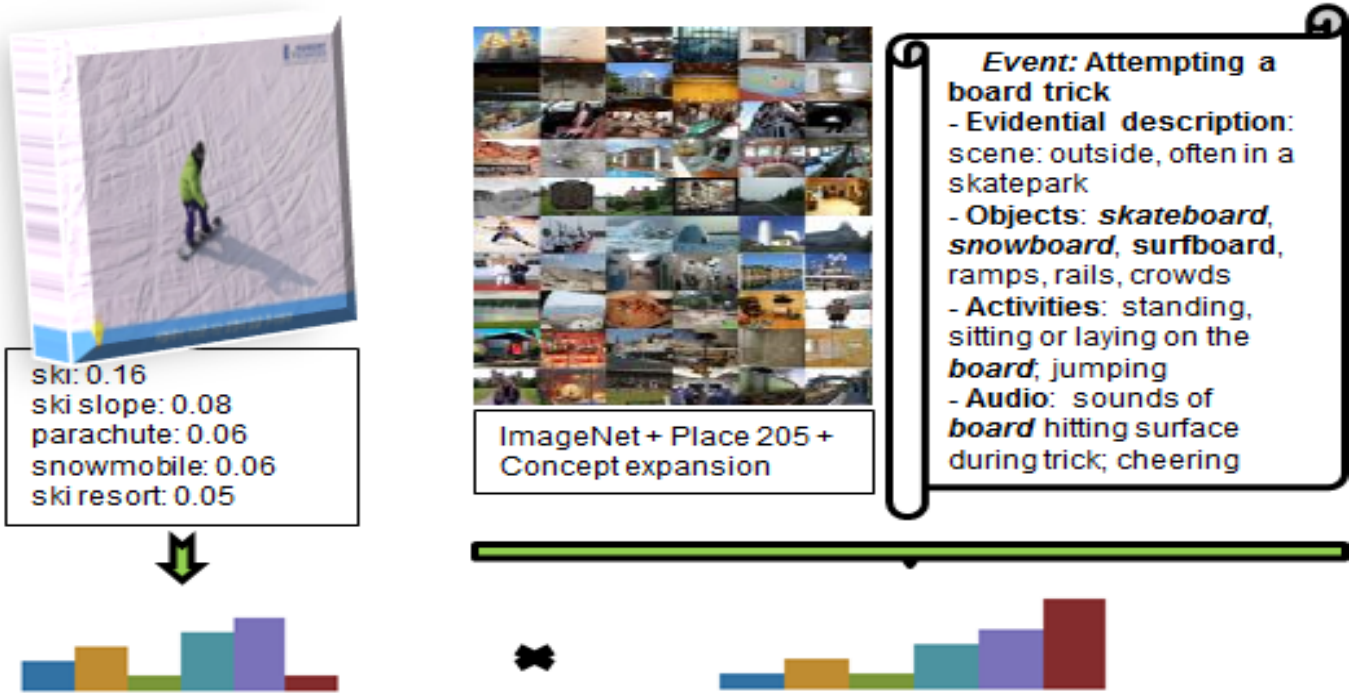
- **Limitation** of previous work: weakly supervised learning problem
  - Instance labels are solely classified based on its represented feature in a large margin framework.
  - The importance of each instance is not considered.
- What if we know information about the instance labels?
  - E.g., the relatedness of each instance to the event of interest.



- **Our proposed method**: A "stronger" supervised learning problem
  - Event-driven Multiple Instance Learning (EDMIL) to utilize the *evidential description* for event detection.

Instance-Event Similarity

- Adopt a **concept expansion** strategy [Wang ICMR2014]
  - Apply at instance level
- consists of 4 steps:
  - Concept detection.
  - Event representation (text-based).
  - Concept-event similarity.
  - Instance-event similarity.



Instance-Event Similarity (cont'd)

Table: Top five concepts discovered by our system for the first 10 events in the MED 2012 dataset.

Event name	Top five importance concepts discovered by our system
Attempting a board trick	Ski, slide rule, ski resort, ski mask, ice skating rink
Feeding an animal	Meat loaf, white shark, food court, pop bottle, cleaver
Landing a fish	Anemone fish, pole, raft, sturgeon, boat deck
Wedding ceremony	Groom, bridegroom, banquet hall, gown, altar
Working on a woodworking project	Jigsaw puzzle, bamboo forest, carpenter's kit, thatch, wooden spoon
Birthday party	Table lamp, lampshade, torch, candle, custard apple
Changing a vehicle tire	Recreational vehicle, car wheel, amphibian, scooter, sports car
Flash mob gathering	Monitor, chime, bell, whistle, ballroom
Getting a vehicle unstuck	Recreational vehicle, amphibian, tank, car wheel, motor scooter
Grooming an animal	Nail, bathtub, shower, fur coat, washbashin

Event-driven Multiple Instance Learning

- V: number of training videos
- I_v: number of instances in video v
- S_{iv}^e: similarity between instance iv and event e
- R: number of level of relatedness from an instance to an event

We define two predict functions for positive and negative instances at level r as follows.

$$P_{pos}(S_{iv}^e, r) = \begin{cases} 1, & \text{if } Rank(S_{iv}^e) \leq r \\ -1, & \text{otherwise} \end{cases}$$

$$P_{neg}(S_{iv}^e, r) = \begin{cases} -1, & \text{if } Rank(S_{iv}^e) \leq r \\ 1, & \text{otherwise} \end{cases}$$

$Rank(.)$ is the function to quantize a similarity into a related level.
Intuition: we can select the top positive and negative instances at each level of relatedness r.

Event-driven Multiple Instance Learning

- Objective function

$$\min_{\{\mathrm{w},\mathrm{b},y,r\}} \frac{1}{2}\|\mathrm{w}\|^2 + \underbrace{C_f \sum_{v=1}^{V}\sum_{i=1}^{I_v} L_f\left(y_{iv},(\mathrm{w}^T\mathrm{x}_{iv}+\mathrm{b})\right)}_{\text{loss due to feature}} + \underbrace{C_p \sum_{v=1}^{V}\sum_{i=1}^{I_v} L_p\left(y_{iv},P(S^e_{iv},r)\right)}_{\text{loss due to prior knowledge}}$$

$$L_f\left(y_{iv},(\mathrm{w}^T\mathrm{x}_{iv}+\mathrm{b})\right) = \max\left(0,1-y_{iv}\left(\mathrm{w}^T\mathrm{x}_{iv}+\mathrm{b}\right)\right)$$

$$L_p\left(y_{iv},P(S_{iv},r)\right) = \begin{cases} S_{iv} & \text{if } P\left(S_{iv},r\right) \neq y_{iv} \\ 0 & \text{otherwise} \end{cases}$$

C_f, C_p: cost parameters to control the influence of each loss function.

- Mixed-integer programming problem $\rightarrow$ **non-convex optimization**!

Optimization Procedure

Alternating optimization strategy to search for a suboptimal solution:

1. Fix instance labels y_{iv} and solve for $\mathbf{w}$ and $b \rightarrow$ classic SVM problem:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C_f \sum_{v=1}^{V} \sum_{i=1}^{I_v} L_f\left(y_{iv}, \mathbf{w}^T \mathbf{x}_{iv} + b\right)$$

2. Fix $\mathbf{w}$ and b, solve for r and update y_{iv}:

$$\min_{y,r} C_f \sum_{v=1}^{V} \sum_{i=1}^{I_v} L_f\left(y_{iv}, \mathbf{w}^T \mathbf{x}_{iv} + b\right) + C_p \sum_{v=1}^{V} \sum_{i=1}^{I_v} L_p\left(y_{iv}, P(S_{iv}^e, r)\right)$$

   Solved by using a **greedy strategy**:
   - Iterate through all level of relatedness to search for the optimal r.
   - Update instance labels y_{iv} using the proposed predict functions.

   Intuition: the most positive and negative instances will be selected first $\rightarrow$ higher possibility to correct mismatched labels.

Experimental Setup

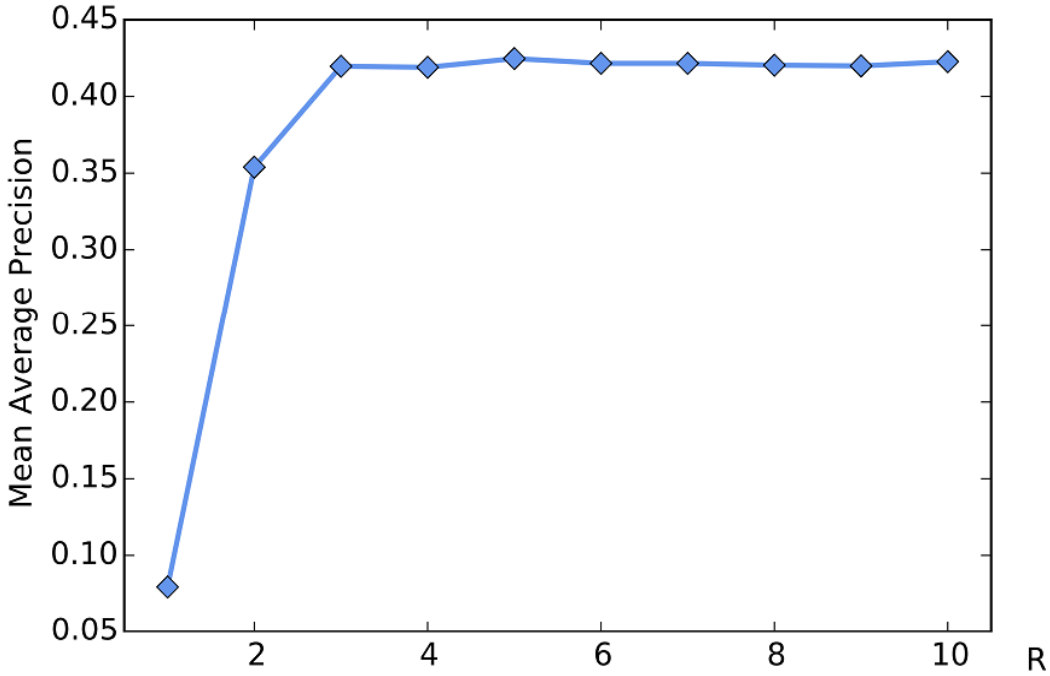
- Dataset

Dataset	No. Event	No. Train Videos	No. Test Videos	Total Videos	Total Hours
MED2010	3	1,744	1,724	3,468	110 hours
MED2011	10	1,331	31,822	33,153	1,100 hours
MED2012	25	3,878	1,938	5,816	250 hours

- Segment length: 8 seconds [Vahdat ICCV2013]
- Feature: Improved Dense Trajectories, MBH descriptor [Wang ICCV2013]
- Feature encoding: Bag-of-words model, 4000 codewords.
- Learning: EDMIL (with linear SVM).
- Testing: Video-level score is obtained by averaging over all instance scores.

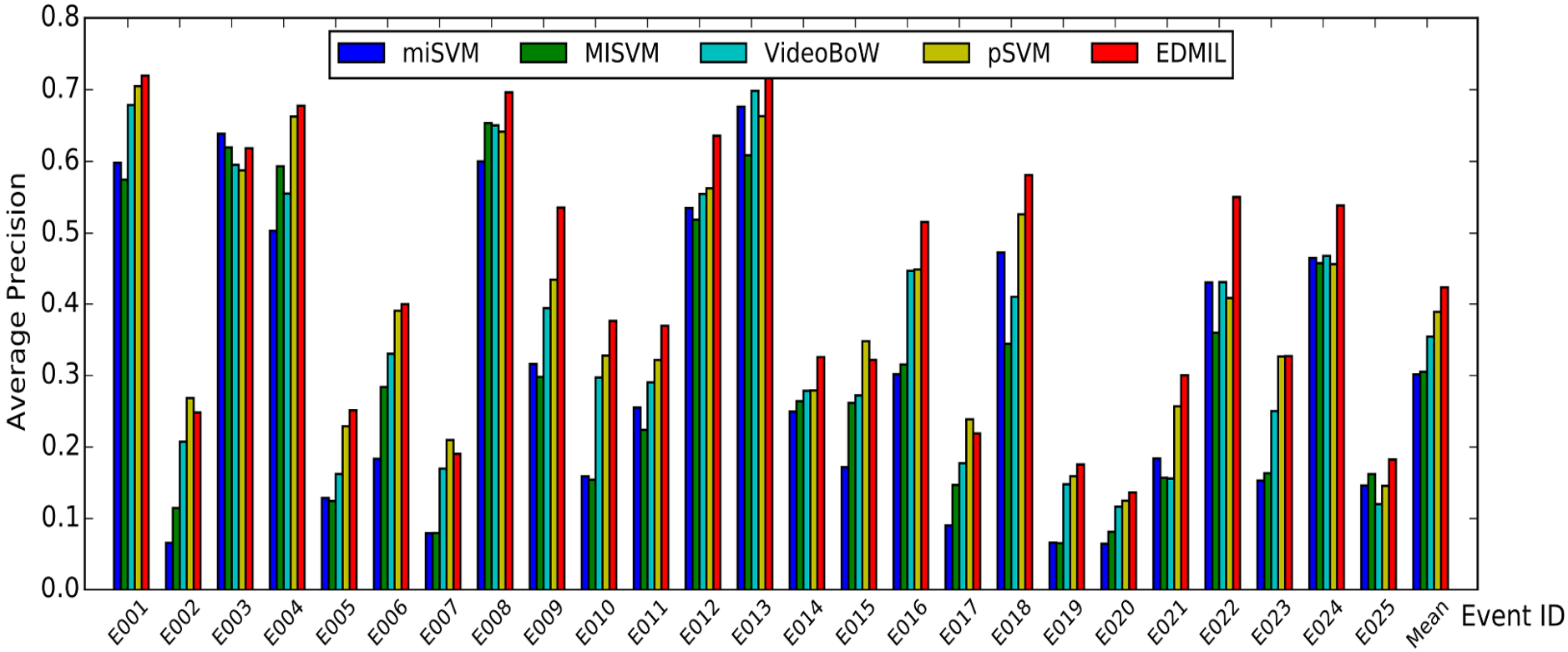
Optimal Number of Related Levels

Select R in the range from 1 to 10 and report the best result.



- Small values of R tend to get low performances $\rightarrow$ the prediction of prior knowledge is not always good, and learning jointly with instance features is necessary.
- The performance becomes saturated when $R > 5 \rightarrow$ we fix the value of R to 5 for further experiments.

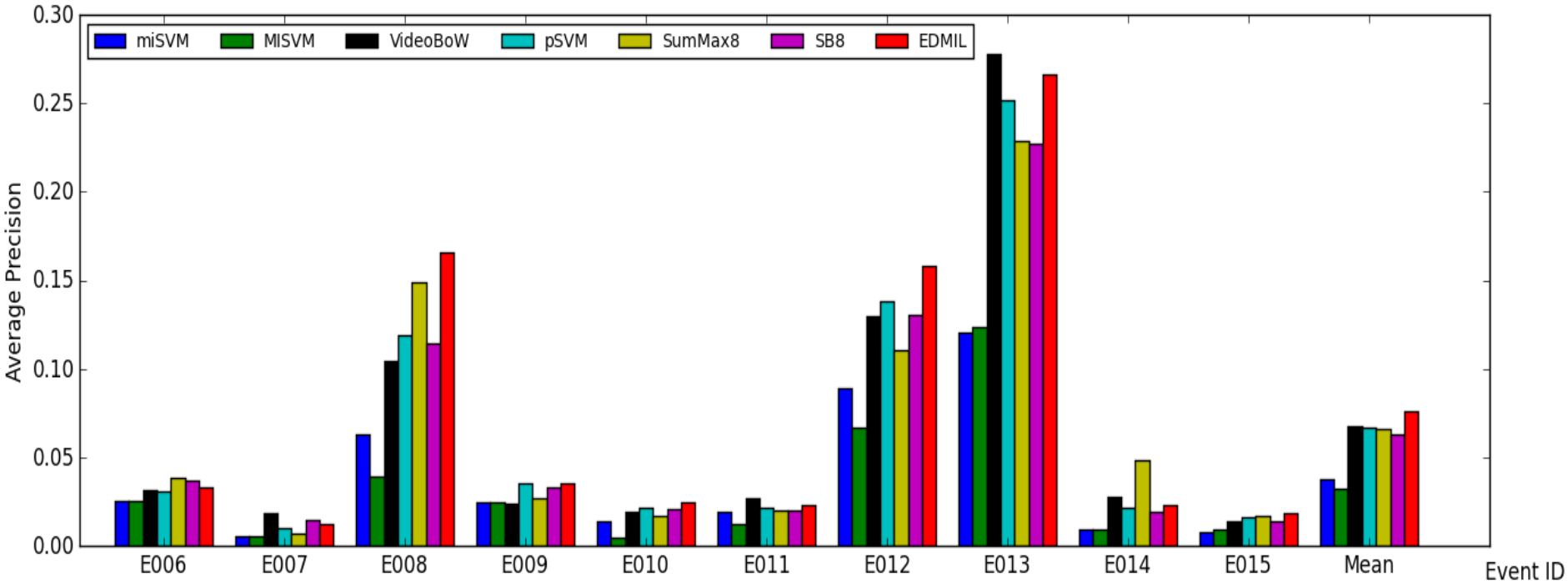
Results on MED2012



The mean APs are 0.3015 (miSVM), 0.3051 (MISVM), 0.3544 (VideoBOW), 0.3890 (pSVM) and **0.4246** (**Ours**).

- Our method significantly outperforms other baselines.
- For the best baseline (pSVM), our method relatively outperforms by 9%.

Results on MED2011



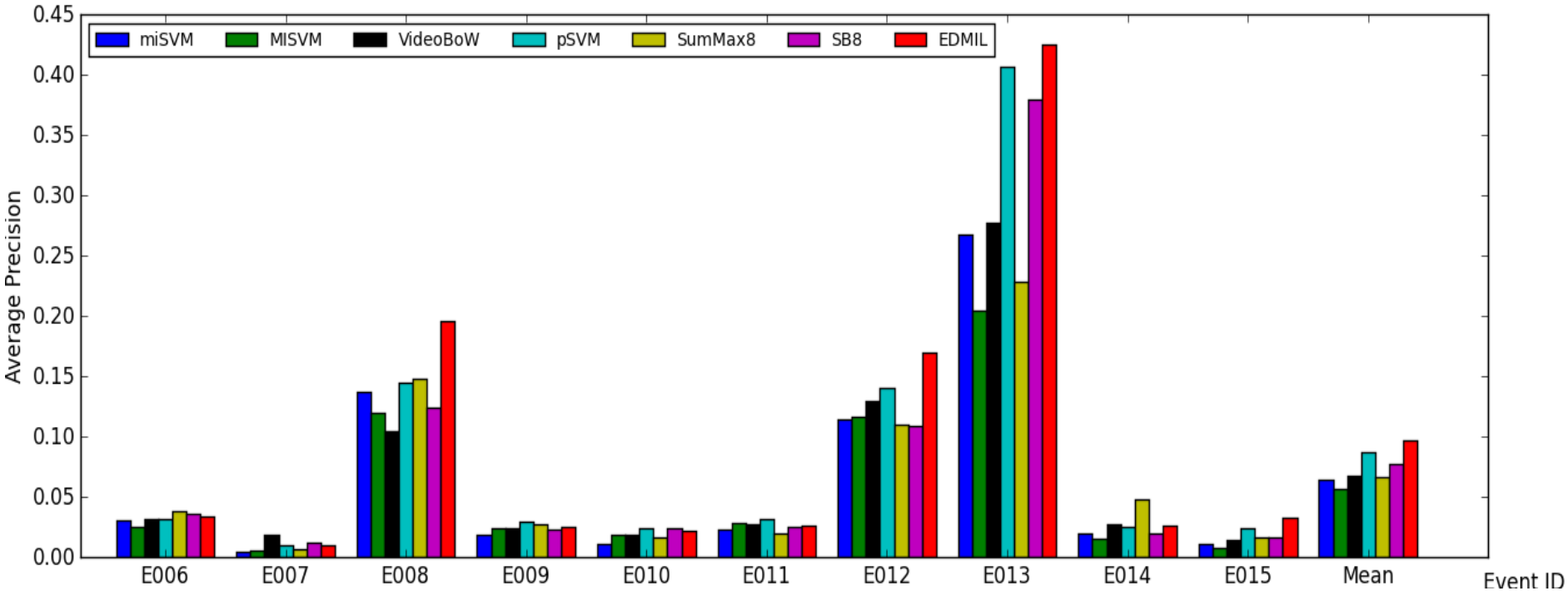
MAP: 0.0378 (miSVM), 0.0322 (MISVM), 0.0674 (VideoBOW), 0.0666 (pSVM), 0.0663 (SM8), 0.0630 (SB8), **0.0761** (**Ours**).

- Our method still performs well (**13%** improvement over the VideoBOW baseline) while pSVM does not.

(At the testing step: Video-level score is obtained by **averaging** its instance scores.)

Results on MED2011

At the testing step: Video-level score is obtained by choosing the **max** instance score.

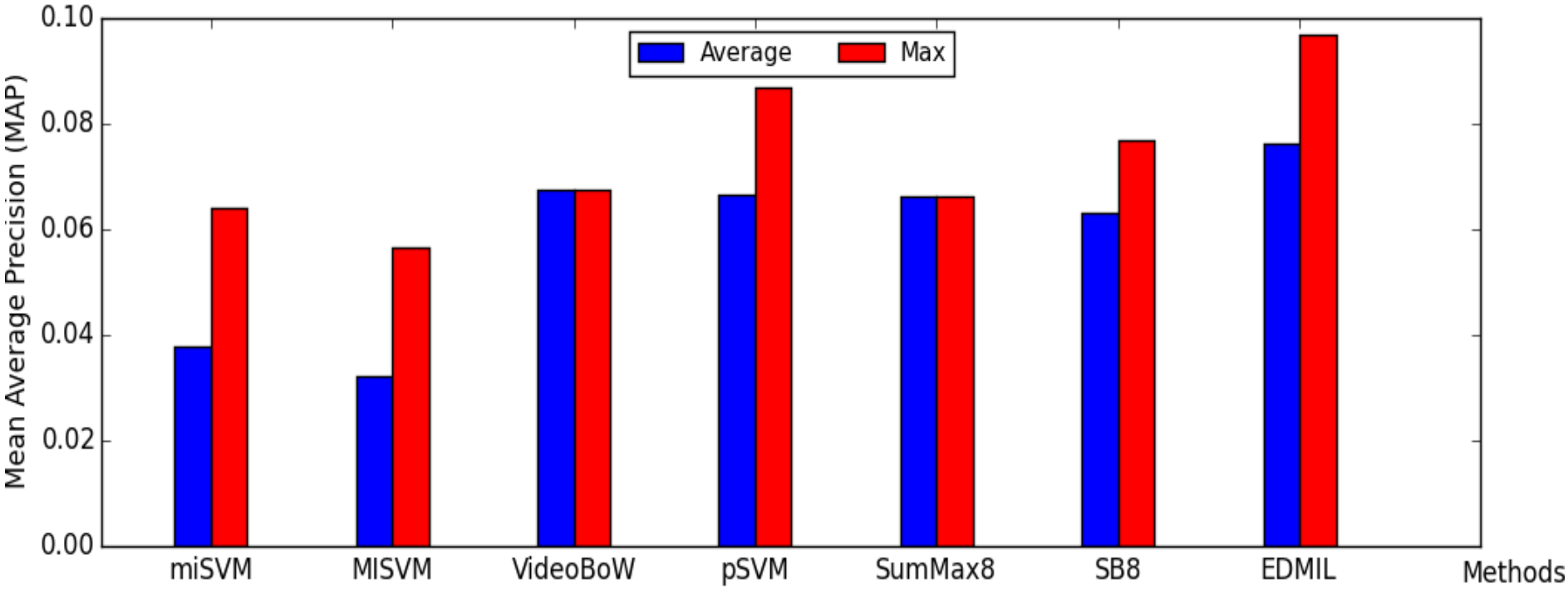


MAP: 0.0640 (miSVM), 0.0564 (MISVM), 0.0674 (VideoBOW), 0.0870 (pSVM), 0.0663 (SM8), 0.0770 (SB8), **0.0968** (**Ours**).

- All methods perform better with this strategy.
- Our method relatively outperforms VideoBOW by **44%**, and pSVM by **11%**.

How to Leverage "*Key Evidences*" for Event Detection?

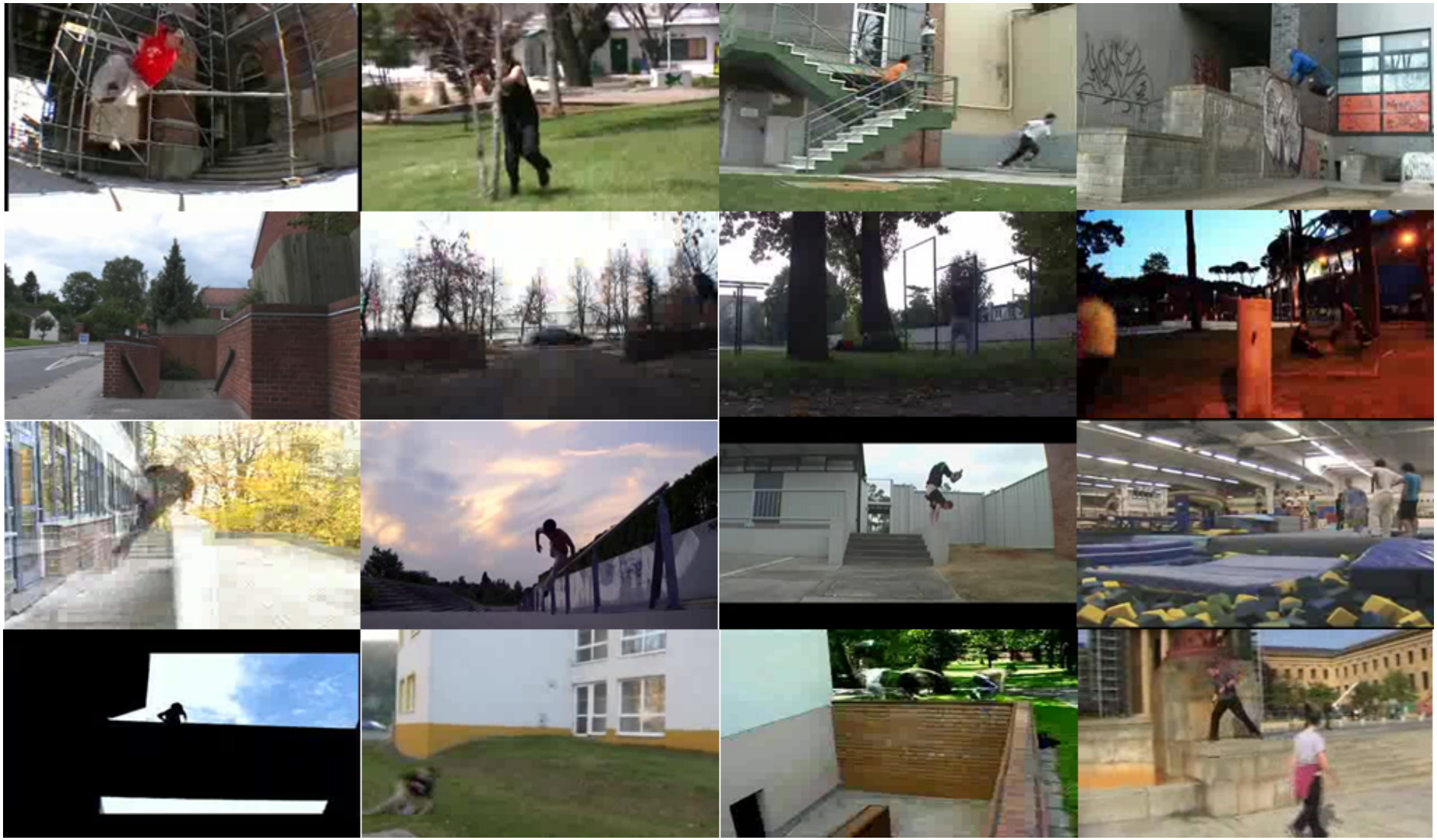
→ At the testing step, video-level score should be obtained by choosing the **max** instance score.



Performance gain when choosing the video score as its **max** instance scores instead of its **average** instance scores.

Top Positive Instances for Event "Parkour"

Parkour: A person travels by foot from one point to another while performing various gymnastic maneuvers.



Top Negative Instances for Event "Parkour"

Parkour: A person travels by foot from one point to another while performing various gymnastic maneuvers.

Conclusions

1. Segment-based Representation (SB)
2. Sum-Max Video Aggregation (SM)
3. **Event-driven Multiple Instance Learning (EDMIL)**
   - A method to leverage the event description to learn key evidences for complex event detection.

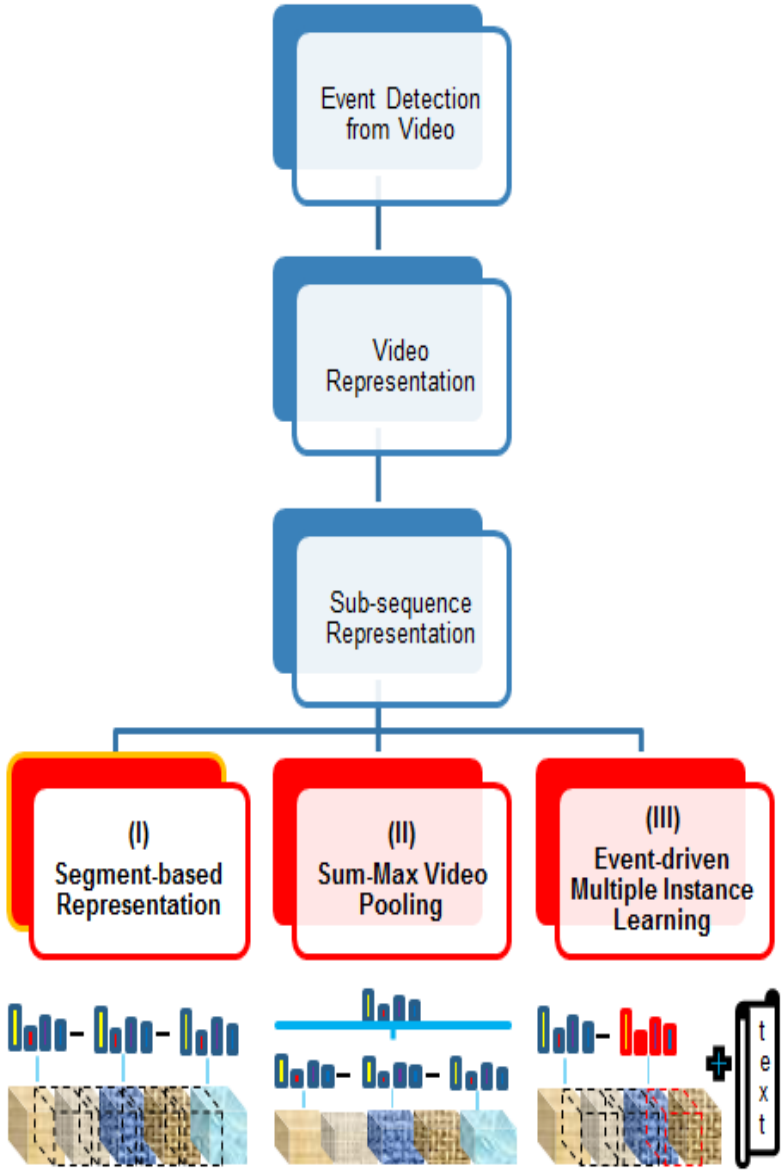
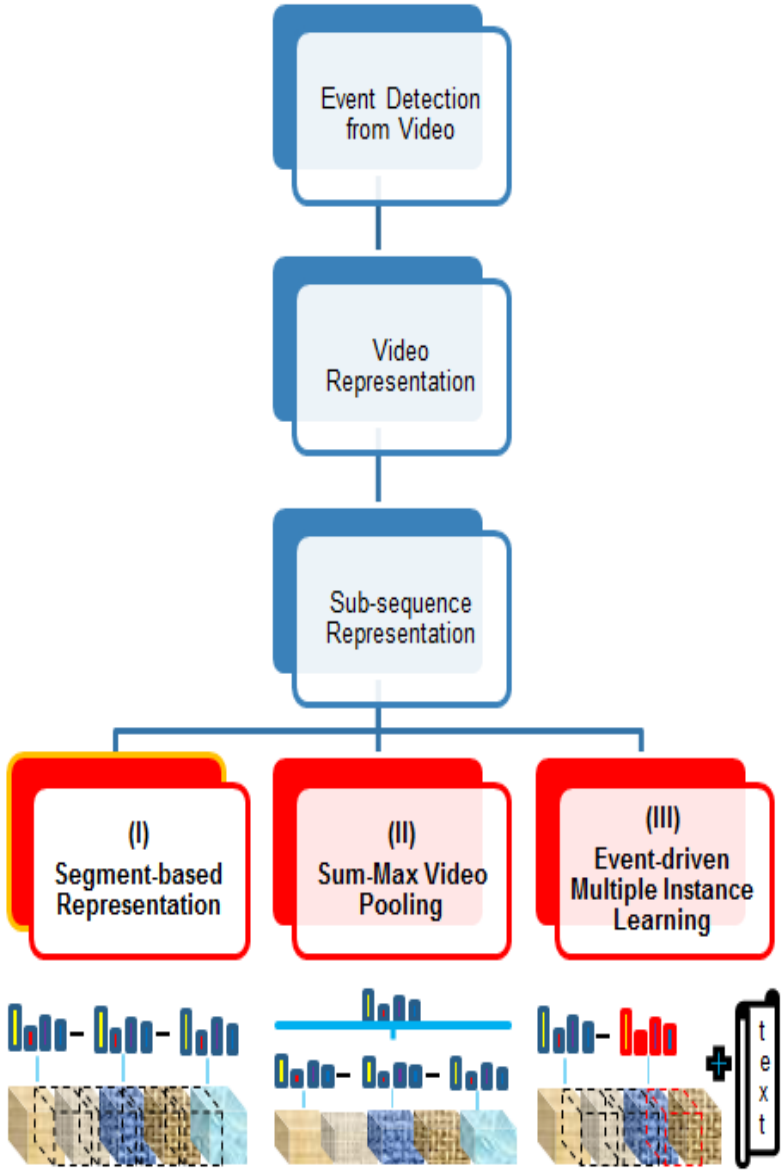


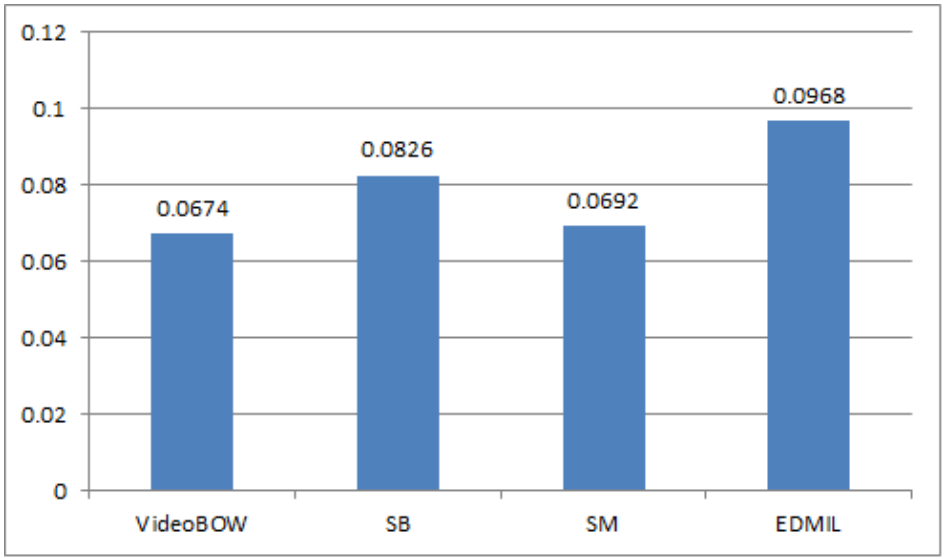
Table of Contents

Summary

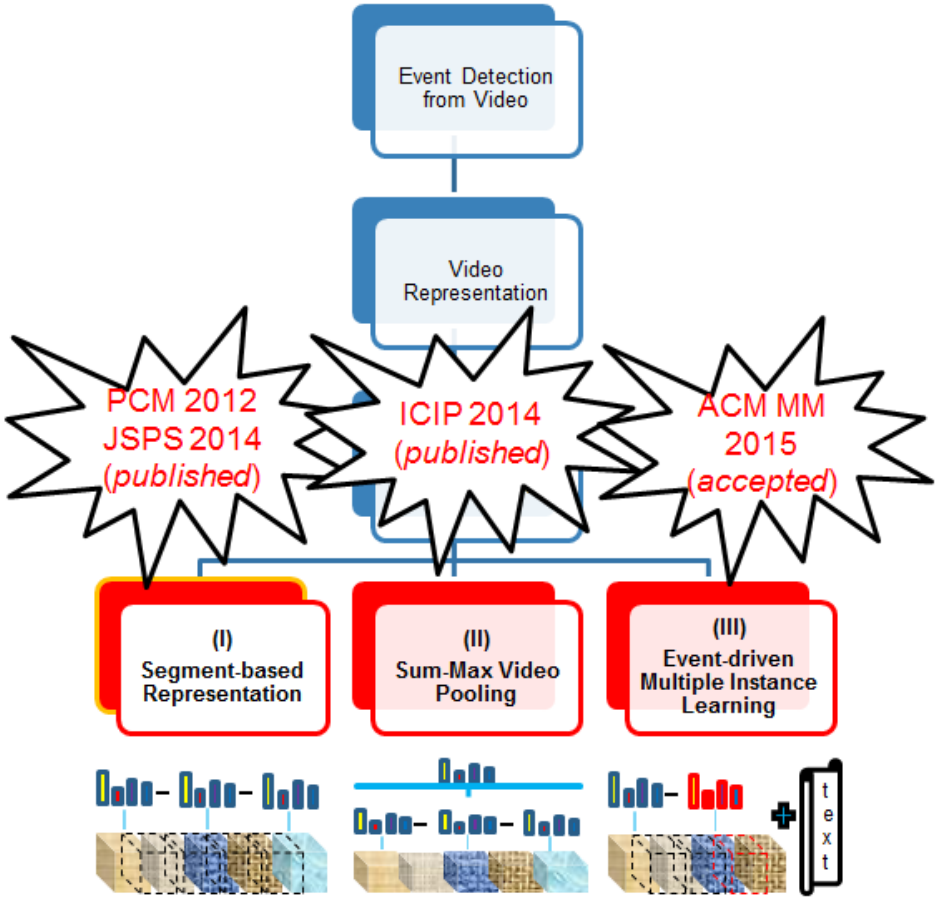
1. Segment-based Representation (SB)
   - Investigate different strategies to decompose a video into segments; study the optimal segment length.
   - *Challenges: uncontrolled capturing, large content variation.*
2. Sum-Max Video Aggregation (SM)
   - An efficient method to aggregate local features into video feature representation.
   - *Challenges: uncontrolled capturing, large scale dataset.*
3. Event-driven Multiple Instance Learning (EDMIL)
   - A method to leverage the event description to learn key evidences for complex event detection.
   - *Challenges: uncontrolled capturing, large content variation.*



Summary (cont'd)



Our methods (SB, SM and EDMIL) improve the baseline VideoBOW by **23%**, **3%** and **44%** respectively on the large scale MED 2011 dataset.



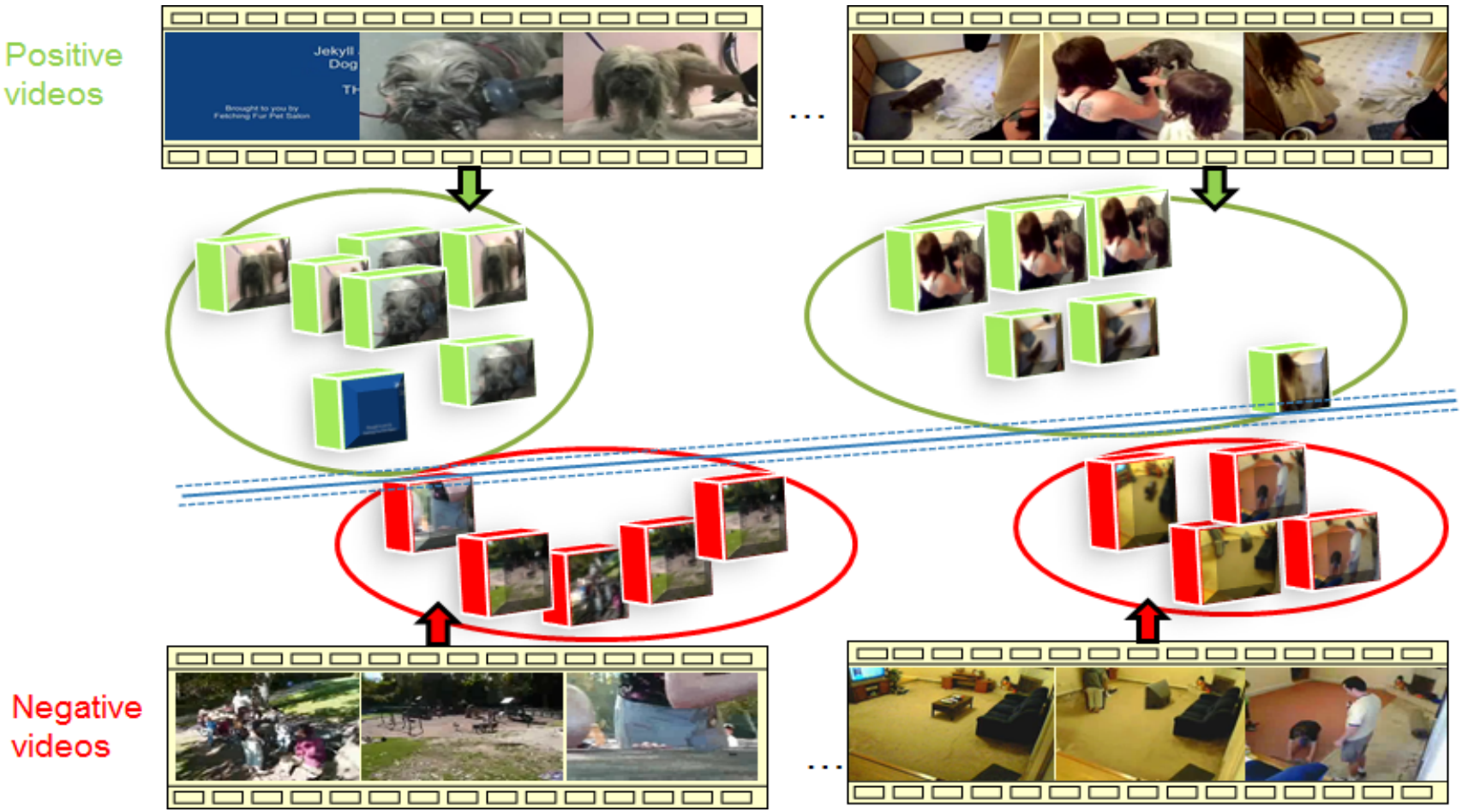
Achievements: PCM2012 *(Rank C)*, ICIP2014 *(Rank B)*, ACMMM2015 *(Rank A)*; JSPS2014 *(IF: 0.6)*.

Thank you for your attention!

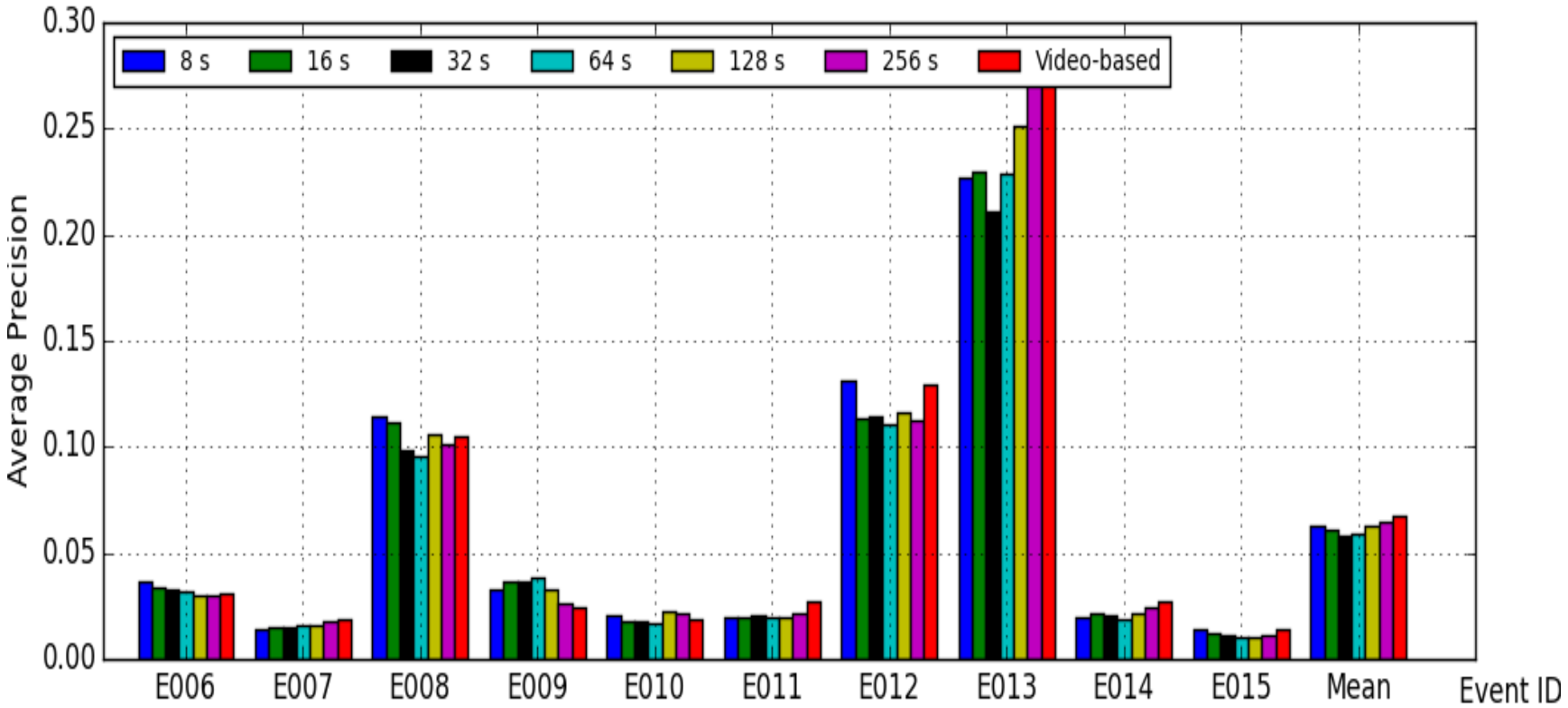
Appendix 1 - Future Work

- Learning the relationship between segments.
  - We have not imposed any constraints on the relationship between segments.
  - However, spatial-temporal relationship might be important to identify an event.
  - For example, in the event "changing a vehicle tire", the action "removing hubcap" should take place before the action "replacing tire".
- Learning the importance of each concept in the concept bank.
  - Concepts are obtained from the event description.
  - We do not know if it really visually represents for that event.
    $\rightarrow$ which concepts that both textually and visually represent for an event?
- Video description generation.
  - Generate textual descriptions for video.
  - Many potential applications such as helping blind people understand what is happening in a video.

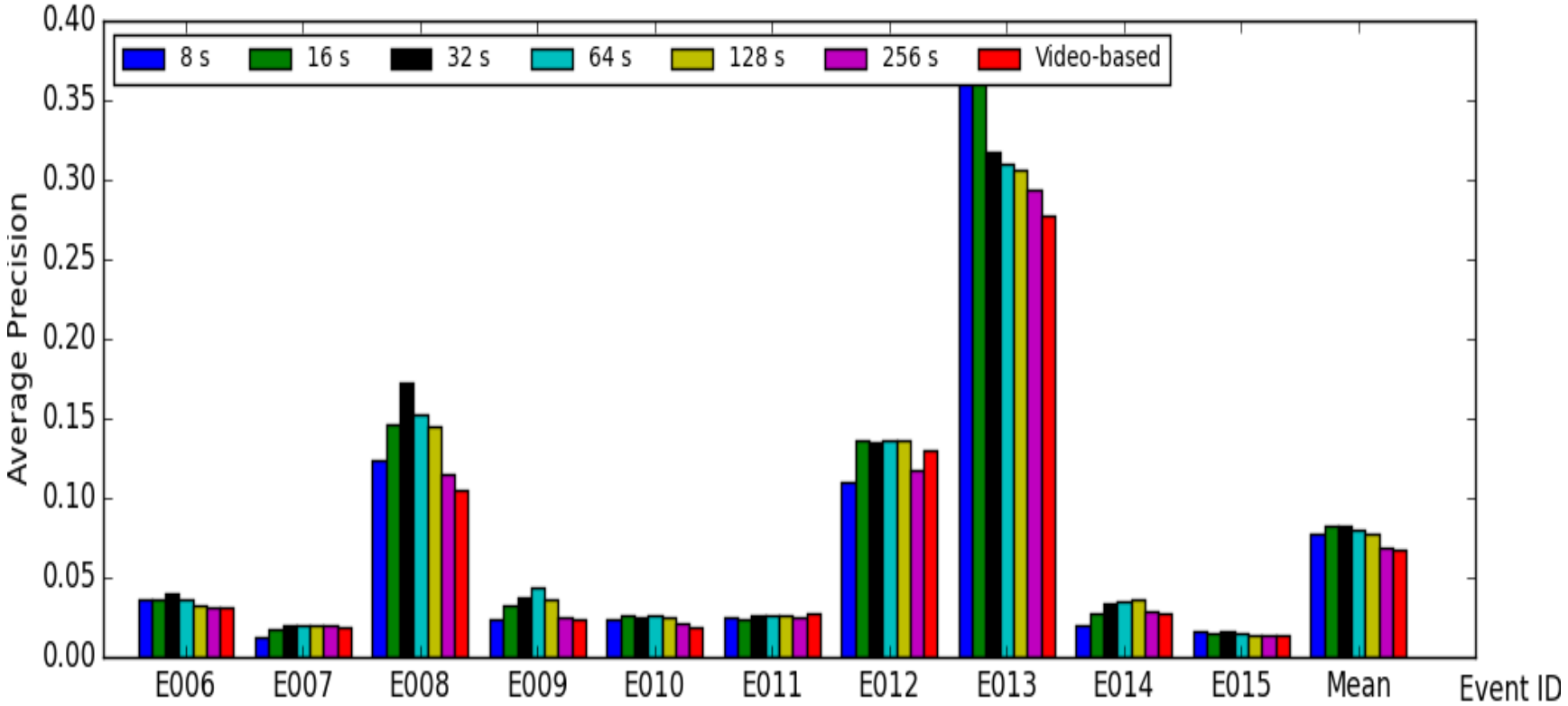
Appendix 2 - Segment-based Representation



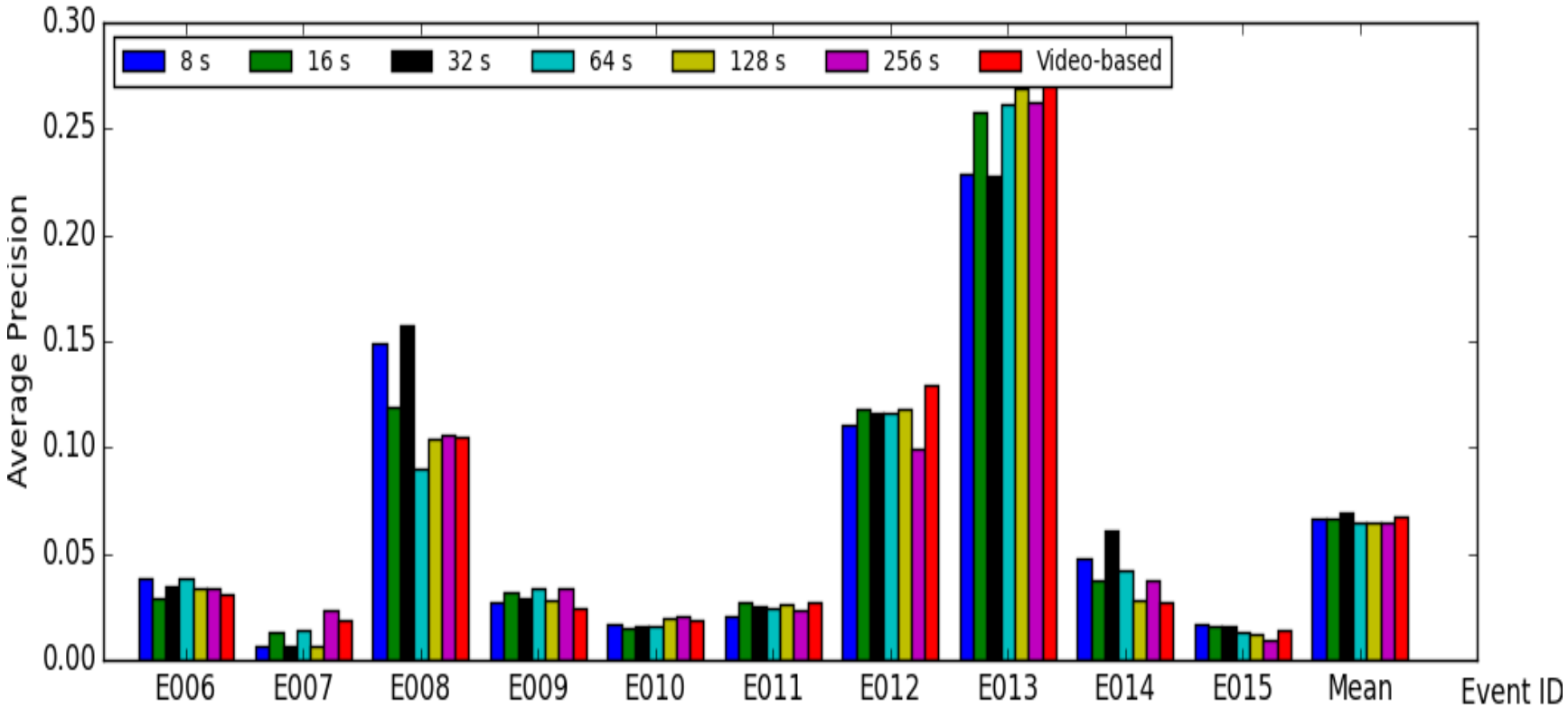
Appendix 3 - Results of Segment-based on MED 2011 (linear SVM, sum aggregation)



Appendix 4 - Results of Segment-based on MED 2011 (linear SVM, max aggregation)



Appendix 5 - Results of Sum-Max video pooling on MED 2011 (linear SVM)



Appendix 6 - Comparation with other reported results

- Dataset: MED11
- Linear SVM
- In terms of mAP (%)

Tang CVPR2011	Cao ECCV2012 (Linear-SAP)	Vahdat ICCV2013 (Linear-LSVM)	Vahdat ICCV2013 (Kernel-LSVM)	Lai CVPR2014 s-pSVM	Lai CVPR2014 m-pSVM	Mori CVPR2015	Ours (EDMIL)
4.77	6.28	3.19	*11.22*	4.3	5.0	6.1	**9.68**