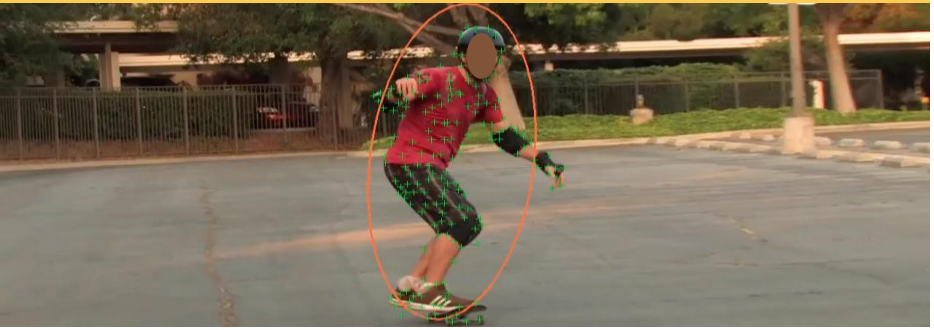
SRI International



SESAME MER

2013 TRECVID Meeting

Bob Bolles

November 21, 2013

SRI International

University of Amsterdam

University of Southern California 1880

Outline

- **MER Demonstration**
- **MED Analysis**
- **MER Analysis**
- **Observations and Future Work**

MER Demonstration – An Example

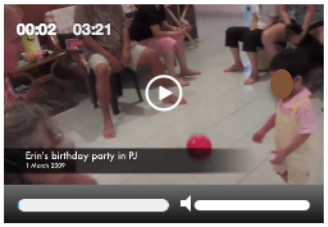
Video Content | Event Query Generation | Event Search | Event Triage | **SESAME**

Event Search | Event Match | Sign Out

Event Search Results

Searching for event **E006 (Birthday Party)** in video data set **MEDTest** from event kit **100ex**

New search

Rank	Video	Observations	Importance	Confidence	Type	Video ID	Event Detection Score
1		Birthday Party for Linda (0:07-0:08)	0.31	0.97	Video OCR	192865	0.759
		Female_Person, Person_clapping, Boy, Male_Person (0:23-0:32)	0.68	0.68	Visual Concepts		
2		birthday party in PJ (0:01-0:05)	0.31	0.98	Video OCR	736791	0.651
		Female_Person, Person_clapping, Person_blowing_candles, Male_Person (2:54-3:04)	0.67	0.67	Visual Concepts		
3		Person_blowing_candles, Boy, Male_Person (2:13-2:17)	0.78	0.78	Visual Concepts	028053	0.638
		Person_clapping, Male_Person, Boy (2:20-2:23)	0.8	0.8	Visual Concepts		
		Person_clapping, Boy, Male_Person (2:55-2:58)	0.81	0.81	Visual Concepts		
		Person_blowing_candles, Male_Person, Female_Person (3:48-3:52)	0.82	0.82	Visual Concepts		
4		Mum's 80th. birthday dinner 27	0.29	0.98	Video	759100	0.626

MED Analysis

Eight Feature- and Concept-based Classifiers

- Visual: 3 classifiers using 1,346 semantic concepts
  - Concepts-HIK *(color histogram analysis)*
  - Concepts-DC *(static image Difference Coding)*
  - SIFT-Fisher *(Fisher encoding of differences)*
- Motion: 2 classifiers
  - DTFV *(Dense Trajectory Fisher Vectors)* and MoSIFT
  - Action Concept HMMFV (96 Sarnoff/UCF actions and UCF 101 actions)
- Audio: 2 classifiers
  - MFCCs *(low-level audio features)*
  - ASR *(Automatic Speech Recognition)*
- Optical Character Recognition (OCR): 1 classifier

Fusion

- Late fusion of the eight results, based on arithmetic mean

Threshold Selection

- Threshold picked to maximize R_0 on a held-out set of data

2013 MED Results

Pre-specified Event Performance

	Visual + Motion	Audio	ASR	OCR	FullSys
100Ex	26.1%	5.9%	4.0%	0.2%	27.6%
10Ex	11.6%	2.6%	1.4%	0.2%	10.3%
0Ex	1.3%		1.7%	2.3%	2.4%

Ad-hoc Event Performance

	Visual + Motion	Audio	ASR	OCR	FullSys
100Ex	23.2%	5.6%	3.9%	0.2%	25.7%
10Ex	12.9%	2.7%	1.4%	0.2%	12.2%
0Ex	1.3%		2.2%	2.2%	2.8%

1. **Our ad hoc performance is essentially the same as pre-specified**
2. **The visual and motion concepts dominate**
3. **Our OCR approach for 0Ex was better than our training-based technique**

MER Analysis

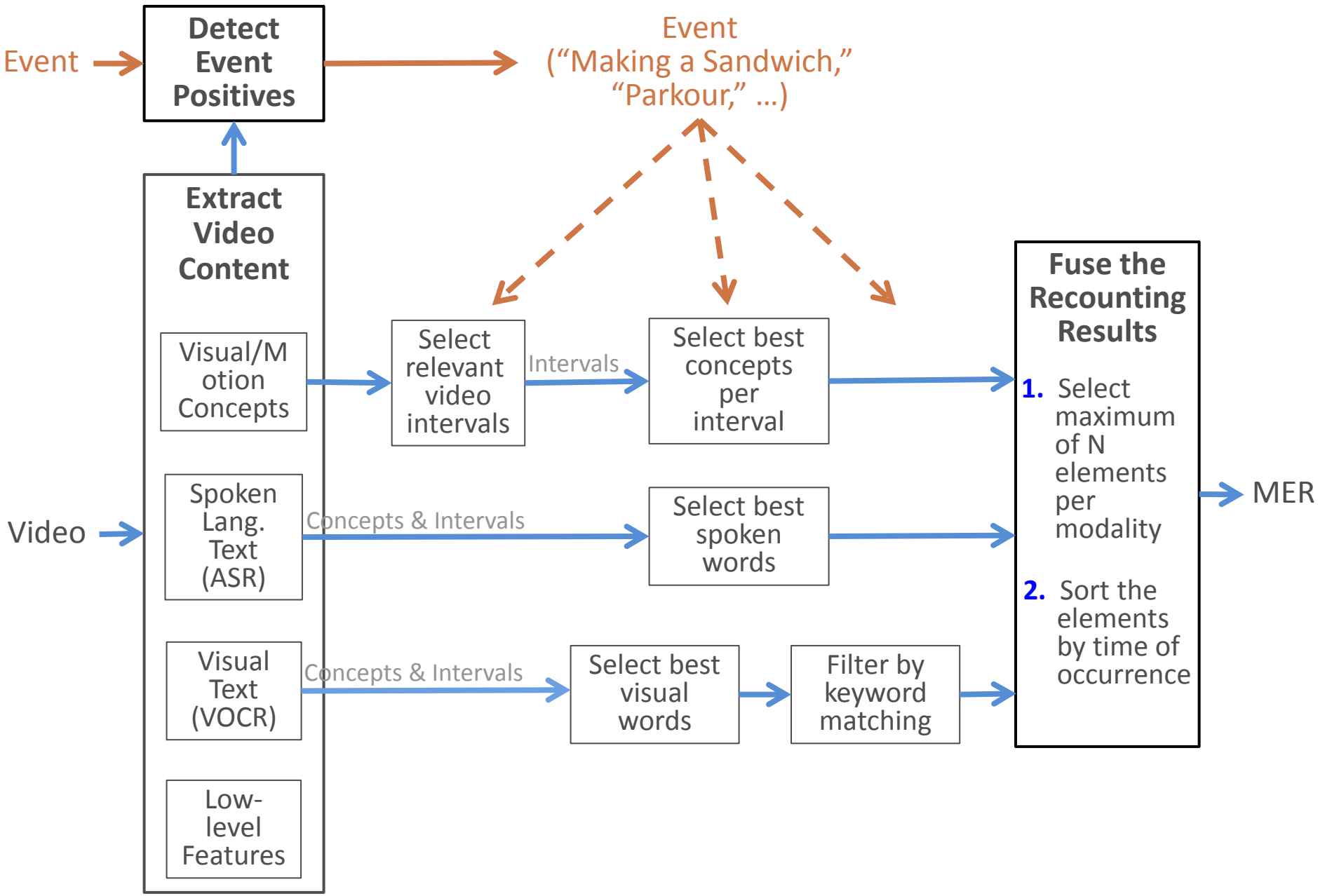
High-level approach

- Each modality (visual, ASR, and OCR) generates a list of their top candidates
- Visual concepts: learn to detect the most discriminative video segments, and then select the most relevant concepts for the event in those segments
- Select a small set of concepts to include in the final list
- Sort (and present) the final list according to their times of occurrence in the video

Used the following to make the final selections

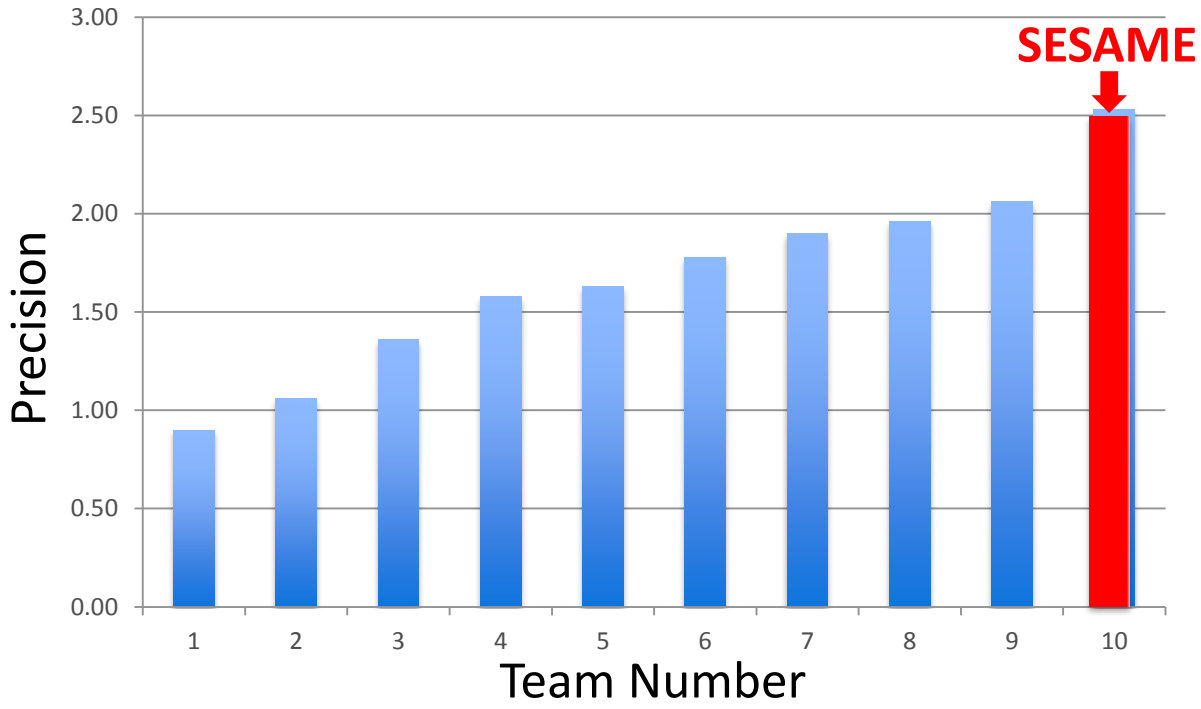
- "Importance" scores, set at training time
- "Confidences" produced by each detector at run time
- Keyword matching of extracted ASR & OCR text to event-specific lists

MER Analysis



MER Results

Accuracy of Judge's final decision: 64.1%
Judge's evaluation of tag quality: 2.53
Percent recounting review time: 41.83%



SESAME achieved the best tag quality

Observations About Our MER Analysis

- **Strategy of identifying key video segments, and then identifying key event-related concepts in those segments worked well**
- **MER contents**
  - Visual concepts in 94% of the videos
  - ASR in 15%
  - OCR in 4%.
- **Our filters on ASR and OCR were too strong** (They eliminated ASR results from 50% of the videos and OCR results from 35%.)
- **For 10Ex and 0Ex, we relied more on substring matching to keyword lists than on importance scores for ASR & OCR**

Future Work

- **Merge overlapping and/or adjacent intervals**
- **Enhance the process that computes the importance of extracted concepts at training time**
- **Develop better normalization of importance scores across visual, action, ASR, and OCR**
- **Enhance the algorithm for automatically generating event-related keywords and their importance scores**

Acknowledgement

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC0067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes nonwithstanding any copyright annotation thereon.