CMU-Informedia @ TRECVID 2013 Multimedia Event Detection

Speaker: Lu Jiang

On behalf of CMU E-LAMP*

Carnegie Mellon University

* Zhen-Zhong Lan, Lu Jiang, Shoou-I Yu, Shourabh Rawat, Yang Cai, Chenqiang Gao, Shicheng Xu, Haoquan Shen, Xuanchong Li, Yipei Wang, Waito Sze, Yan Yan, Zhigang Ma, Wei Tong, Yi Yang, Susanne Burger, Florian Metze, Rita Singh, Bhiksha Raj, Richard Stern, Teruko Mitamura, Eric Nyberg, and Alexander Hauptmann

Acknowledgement

This work was partially supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

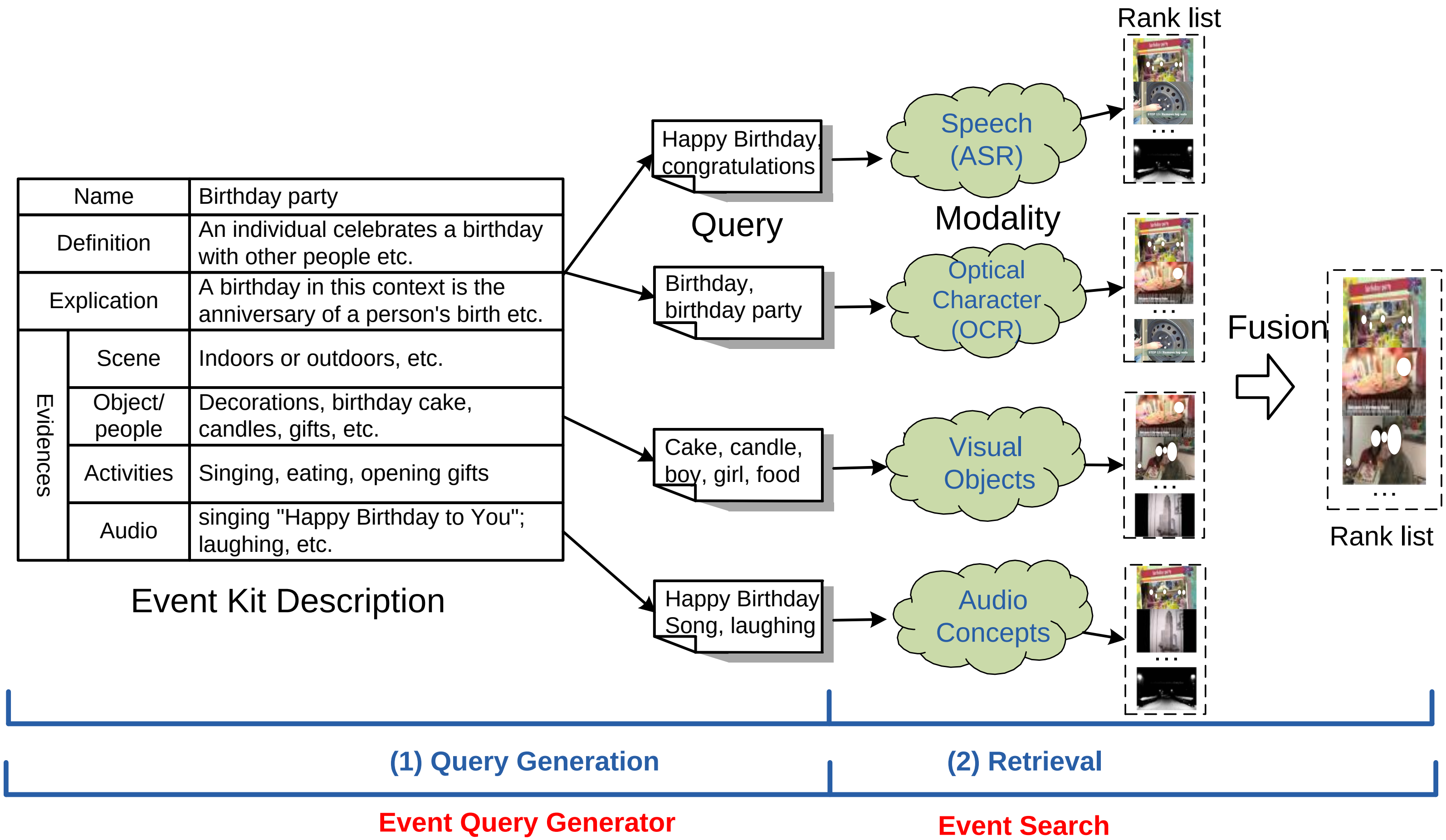
Outline

- MED EK0 Overview
- Related Work
- Pseudo Relevant Set Construction
- Experiment Results
- Conclusions

Outline

- MED EK0 Overview
- Related Work
- Pseudo Relevant Set Construction
- Experiment Results
- Conclusions

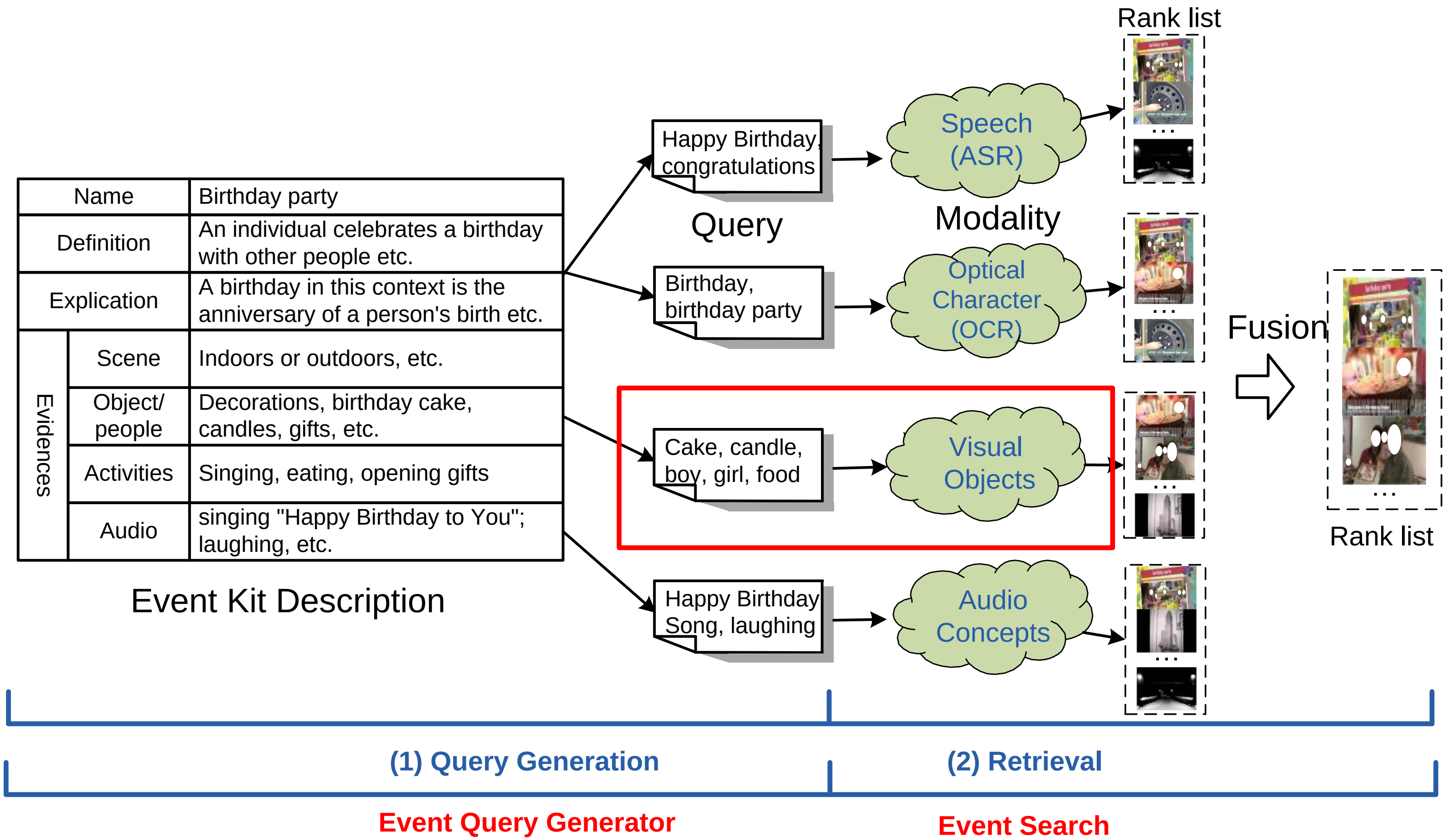
EK0 Pipeline



Non-trivial to use the discriminative low-level features.

Low-level features are non-semantic.

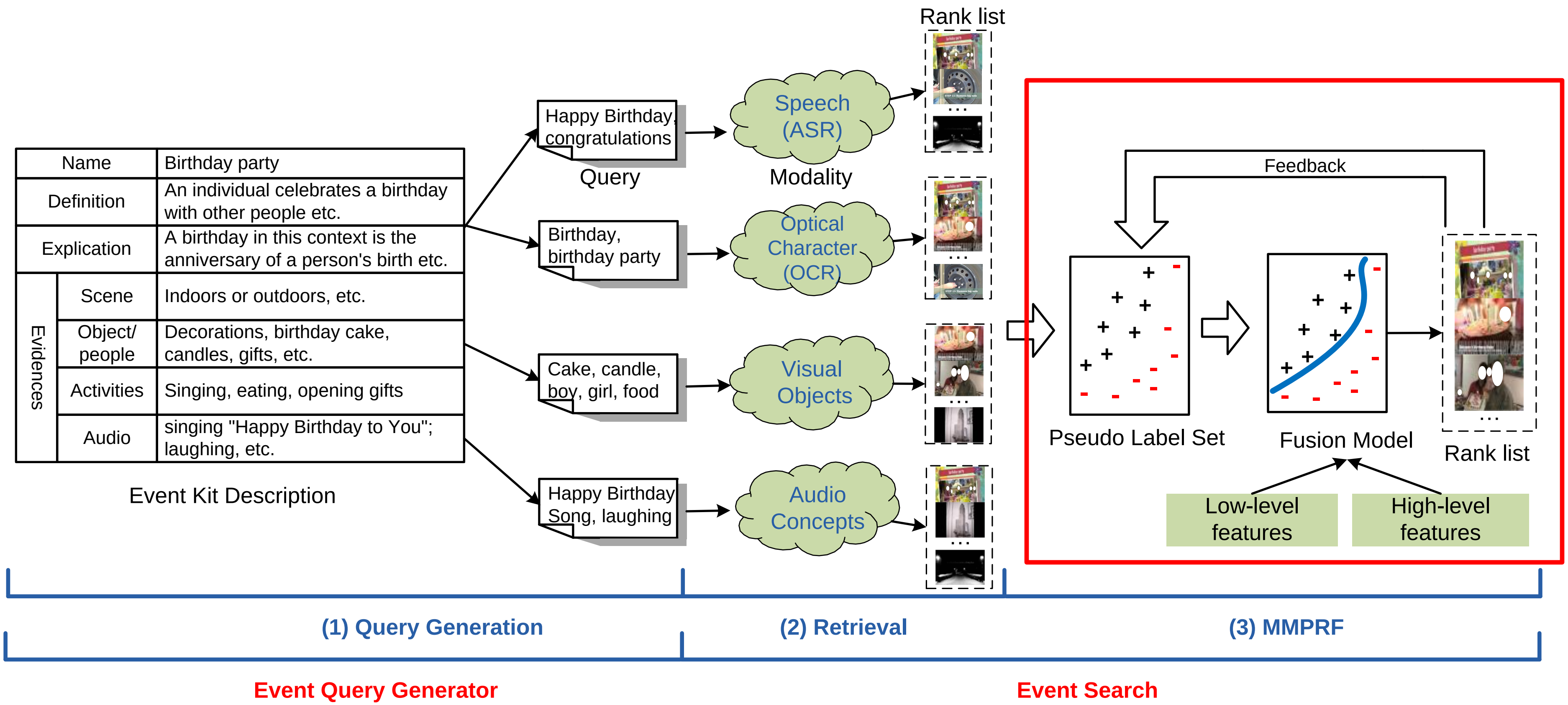
EK0 Pipeline



No way to use the discriminative low-level features such as SIFT.

Low-level features are non-semantic.

EK0 Pipeline with MMPRF



Leverage both high-level features (semantic concepts) and **low-level features**(Dense trajectory and SIFT)

Multi**M**odal **P**seudo **R**elevance **F**eedback (MMPRF)

- MultiModal Pseudo Relevance Feedback (MMPRF) in a nutshell:
  - Construct a pseudo label set.
  - Find a fusion model on the pseudo label set using both high-level and low-level features.
  - Feedback the ranked list of the fusion model to establish the pseudo label set for the next iteration.
- MultiModal: the feedback is carried out on multimodal data (or multiple ranked lists).
- Pseudo: no ground-truth training data or manual relevance judgment is used.
- **MMPRF is completely automatic event search.**

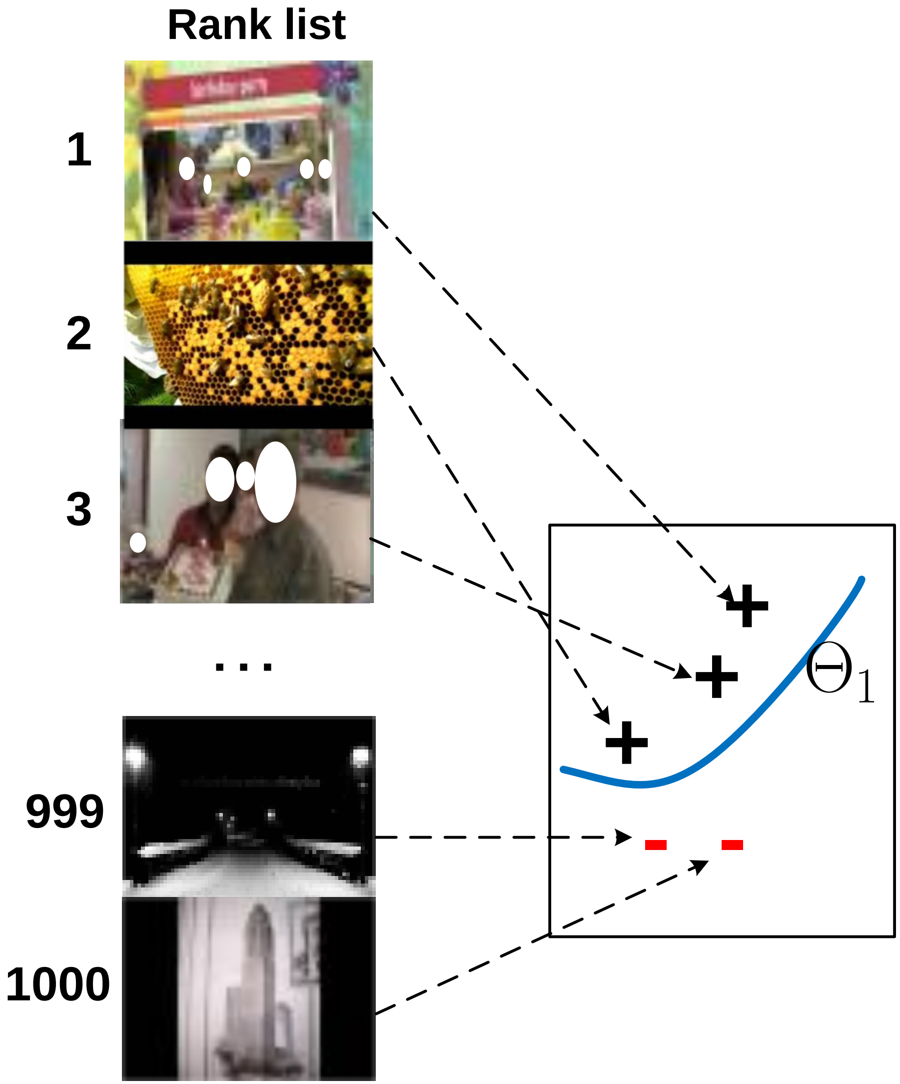
Outline

- MED EK0 Overview
- Related Work
- Pseudo Relevant Set Construction
- Experiment Results
- Conclusions

Pseudo Relevance Feedback(PRF)

- In text retrieval:
  - Rocchio algorithm (Joachims, 1996)
  - Relevance Model (Lavrenko, 2001)
- In multimedia retrieval:
  - Classification-based PRF (Yan, 2003)(Hauptmann, 2008)
  - Learning to rank (Liu, 2008)
- In existing methods, the initial feedback ranking score is obtained from **a single modality**(a single ranked list).

Classification-based PRF (CPRF)



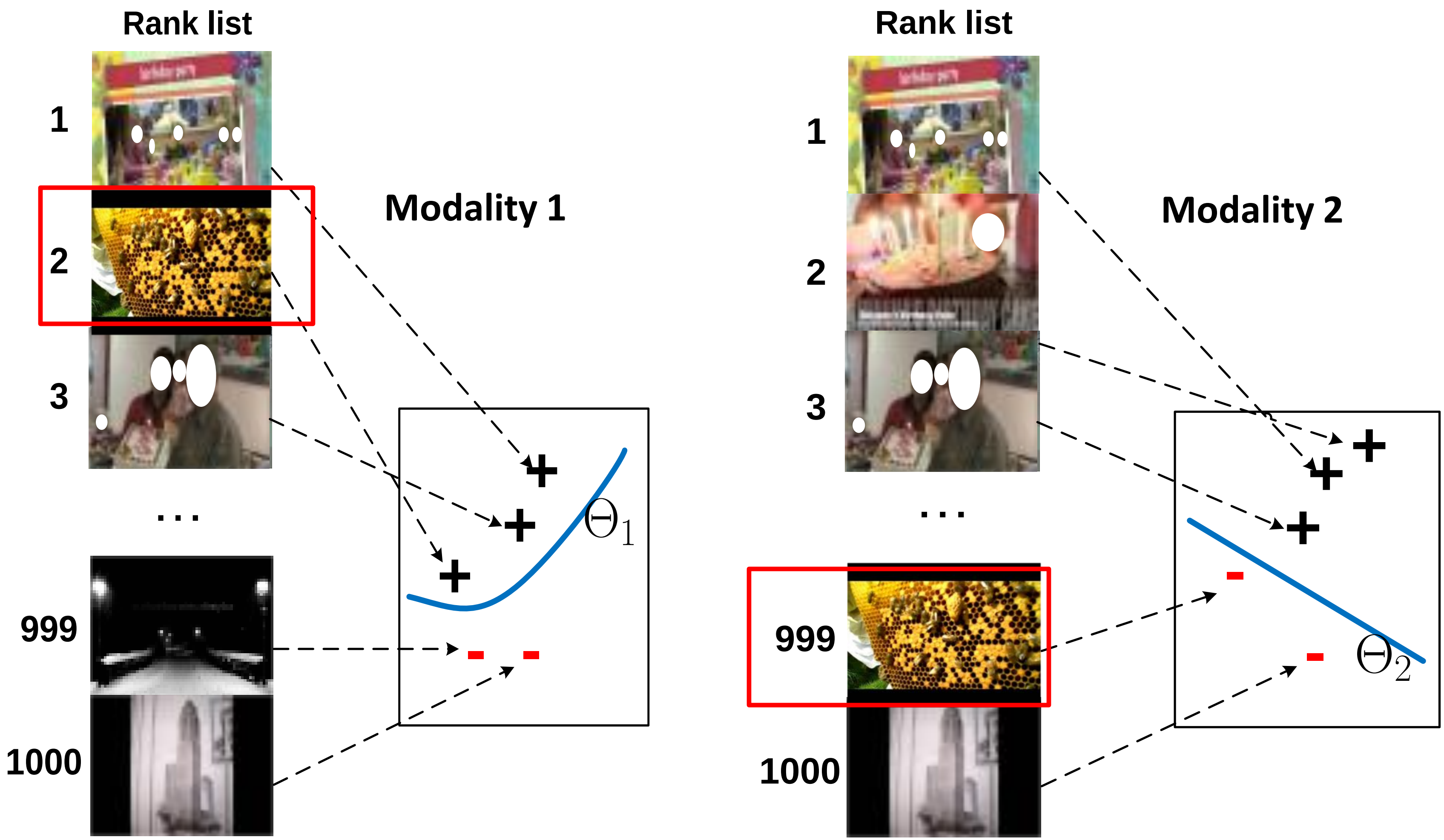
- Treat top-ranked videos as pseudo-positives.
- Treat bottom-ranked videos as pseudo-negatives.

Work reasonably well on unimodal data(Yan, 2003).

Cause inconsistency on multimodal data.

Inconsistency on multimodal data

Considering each modality independently may cause the inconsistency on multimodal data



Outline

- MED EK0 Overview
- Related Work
- Pseudo Relevant Set Construction
- Experiment Results
- Conclusions

Pseudo Relevant Set Construction

- Each modality has its own preference on which pseudo-positives to choose.
- The desired label set satisfies the most modalities.
- The selection process is analogous to voting (**unavailable in the single modality**)
  - Every modality votes for some pseudo-positives.
  - The better the label set fits a modality, the higher the vote is.
  - The set with the highest votes is selected as the pseudo label set.
- A principled approach Maximum Likelihood Estimation (MLE).

MMPRF

- Our objective is to find a pseudo label set that maximizes the likelihood of all modalities.

$$\arg\max_{\mathbf{y}} \sum_{i=1}^{m} \sum_{d_j \in \Omega} y_j \theta_i^T \mathbf{w}_{ij} \text{ s.t.}$$

$$\mathbf{A}^T \mathbf{y} \leq \mathbf{g};\ \mathbf{y} \in \{0,1\}^{|\Omega|}$$

Ω the union of top-ranked videos from all modality;
$\mathbf{y}$ the pseudo label set of the videos in Ω.
θ_i the model parameters of ith modality trained using CPRF.
m the total number of modality.
k^+ the maximum number of pseudo positives to be included in $\mathbf{y}$.
$\mathbf{A}$ the binary matrix $\mathbf{A}_{ij} = 1$ if ith video is in jth modality, 0 otherwise.
$\mathbf{g}$ The modality weight vector, g_i is the number of pseudo-positives to pick from ith modality.

MMPRF

- Our objective is to find a pseudo label set that maximizes the likelihood of all modalities.

$$\boxed{\arg\max_{\mathbf{y}} \sum_{i=1}^{m} \sum_{d_j \in \Omega} y_j \theta_i^T \mathbf{w}_{ij} \ \text{s.t.}}$$

$$\mathbf{A}^T \mathbf{y} \leq \mathbf{g};\ \mathbf{y} \in \{0,1\}^{|\Omega|}$$

Ω the union of top-ranked videos from all modality;
$\mathbf{y}$ the pseudo label set of the videos in Ω.
θ_i the model parameters of ith modality trained using CPRF.
m the total number of modality.
k^+ the maximum number of pseudo positives to be included in $\mathbf{y}$.
$\mathbf{A}$ the binary matrix $\mathbf{A}_{ij} = 1$ if ith video is in jth modality, 0 otherwise.
$\mathbf{g}$ The modality weight vector, g_i is the number of pseudo-positives to pick from ith modality.

- **Objective function: The sum of logarithmic likelihood across all modalities.**

MMPRF

- Our objective is to find a pseudo label set that maximizes the likelihood of all modalities.

$$\arg\max_{\mathbf{y}} \sum_{i=1}^{m} \sum_{d_j \in \Omega} y_j \theta_i^T \mathbf{w}_{ij} \ \text{s.t.}$$

$$\boxed{\mathbf{A}^T\mathbf{y} \leq \mathbf{g};} \mathbf{y} \in \{0,1\}^{|\Omega|}$$

Ω the union of top-ranked videos from all modality;
$\mathbf{y}$ the pseudo label set of the videos in Ω.
θ_i the model parameters of ith modality trained using CPRF.
m the total number of modality.
k^+ the maximum number of pseudo positives to be included in $\mathbf{y}$.
$\mathbf{A}$ the binary matrix $\mathbf{A}_{ij} = 1$ if ith video is in jth modality, 0 otherwise.
$\mathbf{g}$ The modality weight vector, g_i is the number of pseudo-positives to pick from ith modality.

- Objective function: The sum of logarithmic likelihood across all modalities.
- **The constraint controls the maximum number of pseudo-positives to be selected in each modality.**

MMPRF

- Our objective is to find a pseudo label set that maximizes the likelihood of all modalities.

$$\arg\max_{\mathbf{y}} \sum_{i=1}^{m} \sum_{d_j \in \Omega} \boxed{y_j} \theta_i^T \mathbf{w}_{ij} \ \text{s.t.}$$

$$\mathbf{A}^T \mathbf{y} \le \mathbf{g};\ \mathbf{y} \in \{0,1\}^{|\Omega|}$$

Ω the union of top-ranked videos from all modality;
$\mathbf{y}$ the pseudo label set of the videos in Ω.
θ_i the model parameters of ith modality trained using CPRF.
m the total number of modality.
k^+ the maximum number of pseudo positives to be included in $\mathbf{y}$.
$\mathbf{A}$ the binary matrix $\mathbf{A}_{ij} = 1$ if ith video is in jth modality, 0 otherwise.
$\mathbf{g}$ The modality weight vector, g_i is the number of pseudo-positives to pick from ith modality.

- Objective function: The sum of logarithmic likelihood across all modalities.
- The constraint controls the maximum number of pseudo-positives to be selected in each modality.
- **The objective function is linear to the y variable → Integer Programming.**

MMPRF

Carnegie Mellon University

- Our objective is to find a pseudo label set that maximizes the likelihood of all modalities.

$$\arg\max_{\mathbf{y}} \sum_{i=1}^{m} \sum_{d_j \in \Omega} y_j \theta_i^T \mathbf{w}_{ij} \text{ s.t.}$$

$$\mathbf{A}^T \mathbf{y} \leq \mathbf{g}; \boxed{\mathbf{y} \in \{0,1\}^{|\Omega|}}$$

Ω the union of top-ranked videos from all modality;
$\mathbf{y}$ the pseudo label set of the videos in Ω.
θ_i the model parameters of ith modality trained using CPRF.
m the total number of modality.
k^+ the maximum number of pseudo positives to be included in $\mathbf{y}$.
$\mathbf{A}$ the binary matrix $\mathbf{A}_{ij} = 1$ if ith video is in jth modality, 0 otherwise.
$\mathbf{g}$ The modality weight vector, g_i is the number of pseudo-positives to pick from ith modality.

- Objective function: The sum of logarithmic likelihood across all modalities.
- The constraint controls the maximum number of pseudo-positives to be selected in each modality.
- The objective function is linear to the y variable → Integer Programming.
- **Relaxed to linear programming if $\mathbf{0} \leq \mathbf{y} \leq \mathbf{1}$. Efficiently solvable. Time complexity $|\Omega|^3$ (by default $|\Omega| = 50$). Cost less than 10 seconds on a single core on MEDTest.**

Pseudo Label Set Construction With Late Fusion

- Can we use late fusion to construct the pseudo label set? That is first average the scores of all ranked lists and then select the top-ranked videos as pseudo-positives.
- Yes. If we change the objective function.

$$\arg\max_{\mathbf{y}} E[\mathbf{y}|\Omega, \Theta_i] = \sum_{d_j \in \Omega} \boxed{y_j P(y_j|d_j, \Theta_i)} \text{ s.t.}$$

$$\mathbf{J}^T \mathbf{y} \leq k^+ \mathbf{1};\ \mathbf{y} \in \{0, 1\}^{|\Omega|}$$

- Expectation versus Likelihood.
- Late fusion finds the optimal solution to the problem.
- Theoretical justification for the late fusion.
- Unfortunately however, optimizing expectation is **50% worse** than optimizing likelihood on MEDTest.

Modality Weighting

$$\mathbf{A}^T\mathbf{y} \leq \mathbf{g}$$

- At most how many pseudo-positive to select in each modality?
- Estimate using modality accuracy:
  - **Query likelihood**: a modality whose top-ranked videos contain more query words is supposed to be more important.
  - Find indicative words in the event kit description. For example, the occurrence of words “**narration/narrating**” and “**process**” in the event kit description indicates an “accurate ASR event”.

Outline

- MED EK0 Overview
- Related Work
- Pseudo Relevant Set Construction
- Experiment Results
- Conclusions

Results on MEDTest

Events	Method	Single split	Ten splits
Pre-Specified	Without PRF	3.9	4.9 ± 0.8
	Rocchio	5.7	7.4 ± 1.1
	Relevance Model	2.6	3.4 ± 0.5
	CPRF	6.4	8.3 ± 0.9
	Learning to Rank	3.4	4.2 ± 0.7
	MMPRF1	9.0	11.8 ± 1.1
	MMPRF2	**10.1**	**13.6 ± 1.2**
Ad-Hoc	Without PRF	4.0	6.4 ± 0.6
	Rocchio	5.6	6.3 ± 0.9
	Relevance Model	2.3	3.7 ± 0.8
	CPRF	5.9	9.1 ± 1.0
	Learning to Rank	4.3	6.0 ± 0.9
	MMPRF1	7.0	10.9 ± 1.0
	MMPRF2	**8.3**	**12.1 ± 1.1**

- MMPRF1: w/o modality weighting. MMPRF2: w/ modality weighting.
- Improve the baseline Without PRF by a **relative 158%** (**absolute 6.2%**) on Pre-Specified events and by a **relative 107%** (**absolute 4.3%**) on Ad-Hoc events.
- **Statistically significantly better** than other baseline methods.

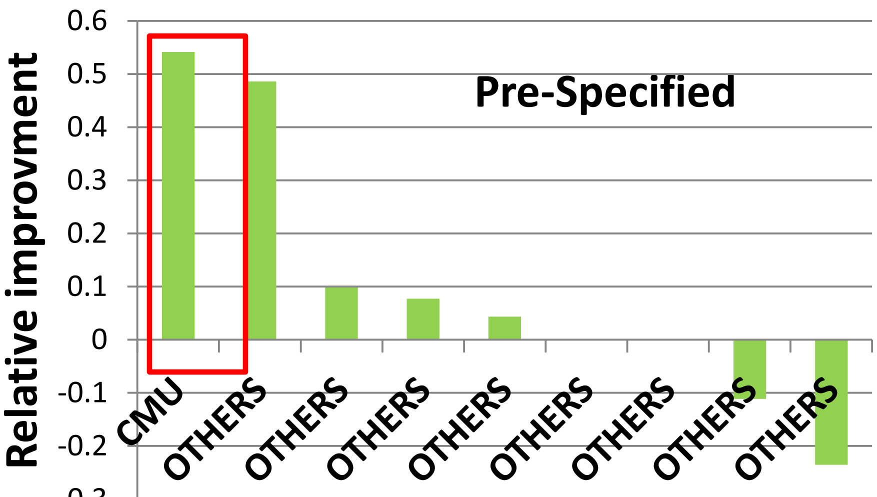
Official Results on PROGAll

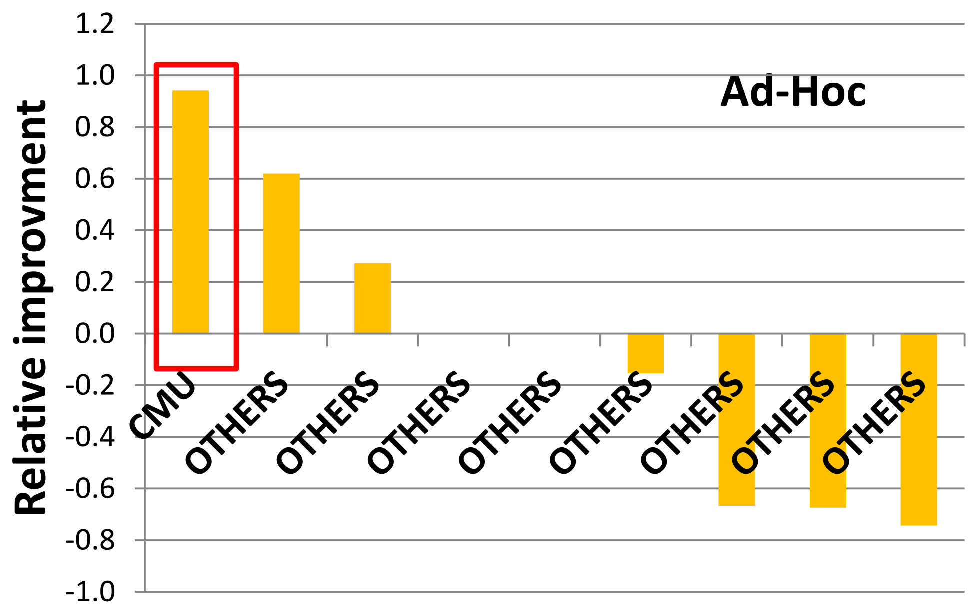
EK0 Results	FullSys	ASRSys	AudioSys	OCRSys	VisualSys
CMU Pre-Specified	3.7	1.8	0.3	2.1	2.4
CMU Ad-Hoc	10.1	3.1	0.2	2.8	5.2

- MMPRF is only used in our FullSys.
- Relative improvement of the FullSys over the best sub-system.
  - by a **relative 54%** on Pre-Specified events.
  - by a **relative 94%** on Ad-Hoc events.

Results on PROGAll

- Relative Improvement of FullSys over the best SubSys.





Outline

- MED EK0 Overview
- Related Work
- Pseudo Relevant Set Construction
- Experiment Results
- Conclusions

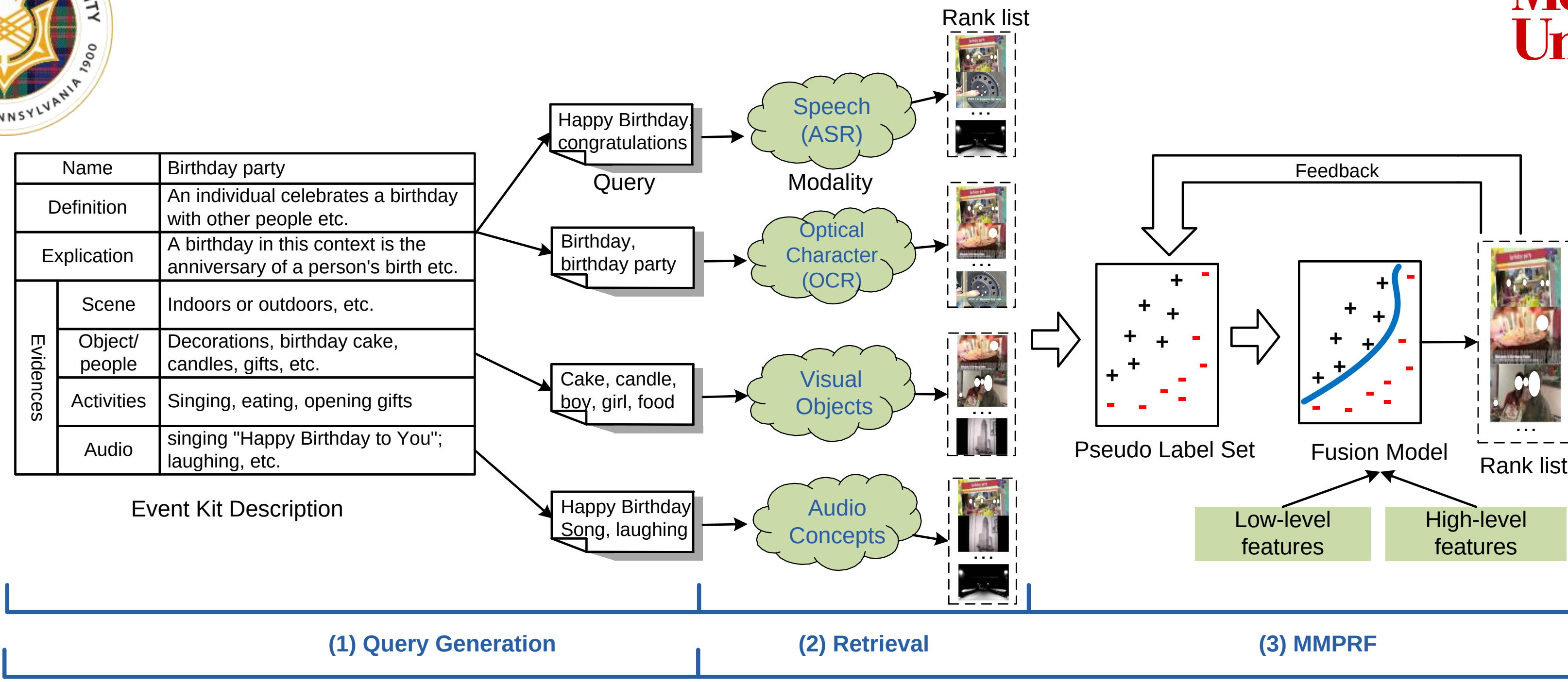
Conclusions

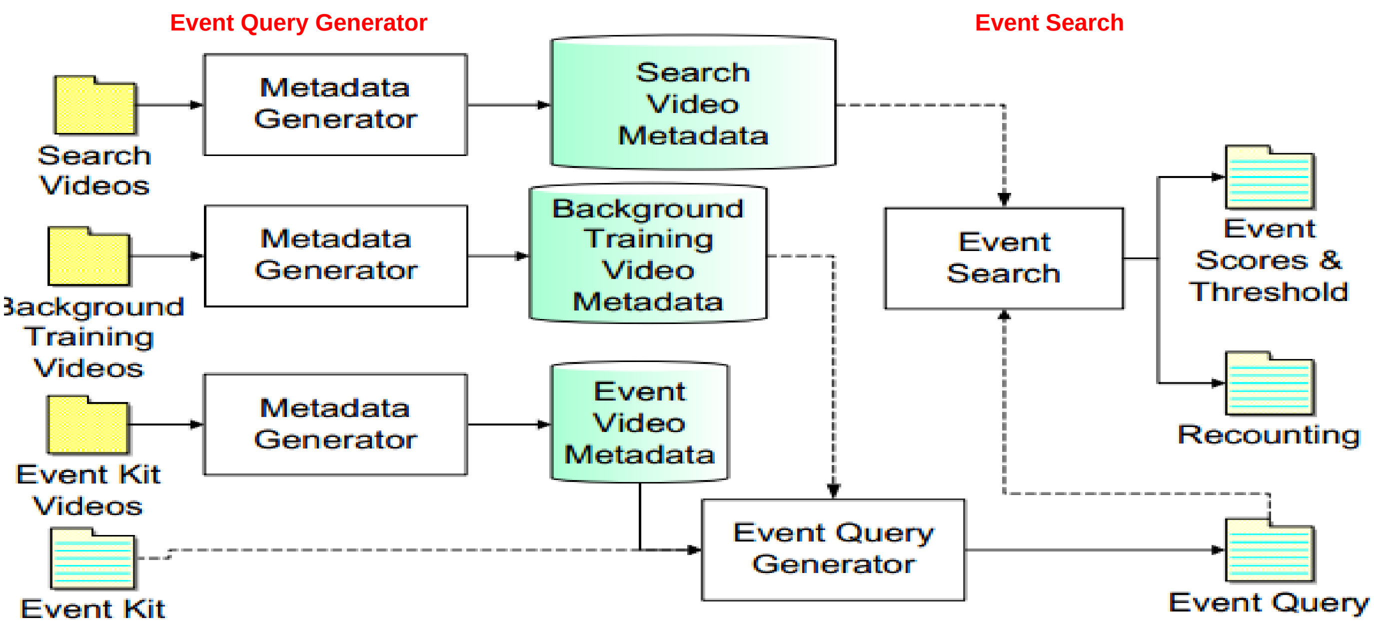
- A few things to take away from this talk:
  - MultiModal Pseudo Relevance Feedback (MMPRF) is a first attempt to use both high-level and low-level features in MED EK0.
  - MMPRF offers a solution to conduct PRF on multiple ranked lists. Empirically it significantly outperforms all baseline methods on MEDTest.
  - Modality weighting is beneficial.

References

- Joachims, Thorsten. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*. No. CMU-CS-96-118. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE, 1996.
- Yan, Rong, Alexander G. Hauptmann, and Rong Jin. "Negative pseudo-relevance feedback in content-based video retrieval." *Proceedings of the eleventh ACM international conference on Multimedia*. ACM, 2003.
- Lavrenko, Victor, and W. Bruce Croft. "Relevance based language models."*Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001.
- Liu, Yuan, et al. "Learning to video search rerank via pseudo preference feedback." *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008.
- Bao, Lei, et al. "Informedia@ trecvid 2011." *TRECVID2011, NIST* (2011).
- Jiang, Lu, Alexander G. Hauptmann, and Guang Xiang. "Leveraging high-level and low-level features for multimedia event detection." *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012.
- Tong, Wei, et al. "E-LAMP: integration of innovative ideas for multimedia event detection." *Machine Vision and Applications* (2013): 1-11.
- Hauptmann, Alexander G., Michael G. Christel, and Rong Yan. "Video retrieval based on semantic concepts." *Proceedings of the IEEE* 96.4 (2008): 602-622.

THANK YOU.
Q&A?


Name
Birthday party
Definition
An individual celebrates a birthday with other people etc.
Explication
A birthday in this context is the anniversary of a person's birth etc.
Evidences
Scene
Indoors or outdoors, etc.
Object/ people
Decorations, birthday cake, candles, gifts, etc.
Activities
Singing, eating, opening gifts
Audio
singing "Happy Birthday to You"; laughing, etc.
Event Kit Description
Happy Birthday, congratulations
Query
Birthday, birthday party
Cake, candle, boy, girl, food
Happy Birthday Song, laughing
Speech (ASR)
Modality
Optical Character (OCR)
Visual Objects
Audio Concepts
Rank list
Feedback
Pseudo Label Set
Fusion Model
Rank list
Low-level features
High-level features
(1) Query Generation
(2) Retrieval
(3) MMPRF


Event Query Generator
Event Search
Search Videos
Metadata Generator
Search Video Metadata
Background Training Videos
Metadata Generator
Background Training Video Metadata
Event Kit Videos
Metadata Generator
Event Video Metadata
Event Kit
Event Query Generator
Event Search
Event Scores & Threshold
Recounting
Event Query