Semantic Search in Internet Videos

Lu Jiang, Deyu Meng, Shoou-I Yu, Shiguang Shan, Teruko Mitamura and Alex Hauptmann

Carnegie Mellon University

Carnegie Mellon University

People

- CMU Informedia Team

Deyu Meng

Shoou-I Yu

Shiguang Shan

Teruko Mitamura

Alexander Hauptmann

Carnegie Mellon University

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.

Carnegie Mellon University

Outline

- Introduction
- Background
- Breakthrough 1: Large-scale Semantic Indexing (SIN)
- Breakthrough 2: Self-Paced Reranking (SPaR)
- Conclusions

Carnegie Mellon University

Outline

- **Introduction**
- Background
- Breakthrough 1: Large-scale Semantic Indexing (SIN)
- Breakthrough 2: Self-Paced Reranking (SPaR)
- Conclusions

Carnegie Mellon University

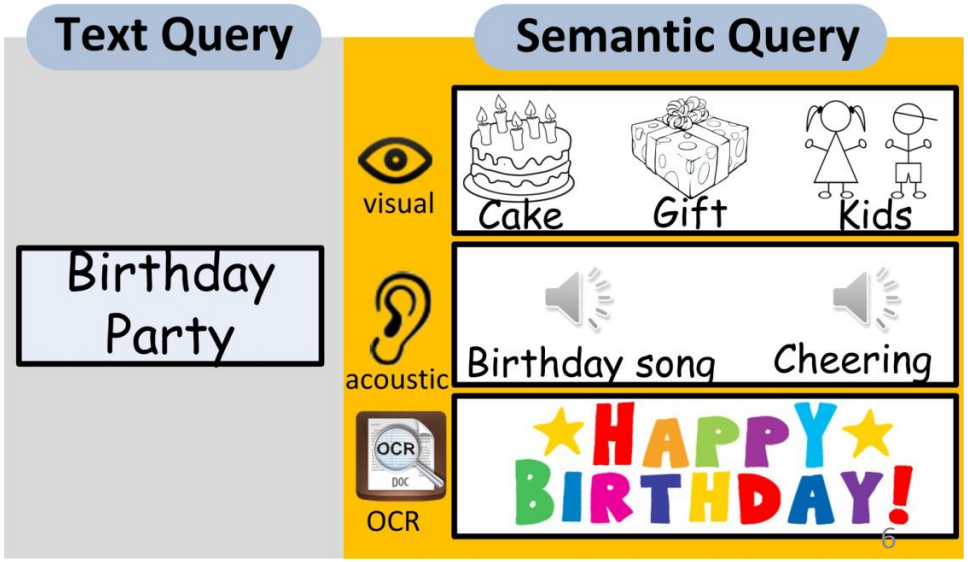
Semantic Search in Internet Videos

Metadata is insufficient in describing the content of a video:

- Biased views/Irreverent to the content
- Weak temporal describability
- Can be missing

Next generation **semantic search engine** for Internet video (called **Zero-Example Search** or **0Ex**):

- No textual metadata
- No example videos
- Complex content query
- Multimodal query



Carnegie Mellon University

Zero-Example Search

- Event of Interest: Birthday Party

ELAMP
Prototype System

Carnegie Mellon University

Zero-Example Search

- Event of Interest: Birthday Party

ELAMP
Prototype System

Birthday Cake, Kids, Gift

Carnegie Mellon University

Zero-Example Search

- Event of Interest: Birthday Party

ELAMP
Prototype System

Birthday Cake, Kids, Gift

Happy Birthday, Cheering



Carnegie Mellon University

Zero-Example Search

- Event of Interest: Birthday Party

ELAMP
Prototype System

Birthday Cake, Kids, Gift

Happy Birthday, Cheering

Birthday Party

Carnegie Mellon University

Zero-Example Search

- Event of Interest: Birthday Party

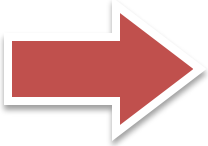
ELAMP
Prototype System

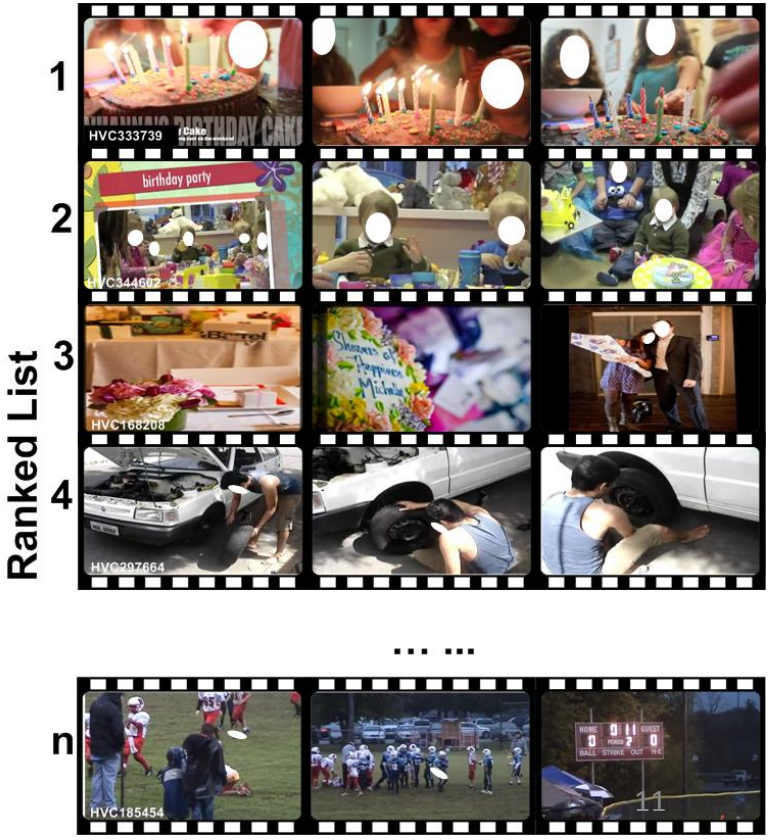
Birthday Cake, Kids, Gift

Happy Birthday, Cheering

Birthday Party

0Ex System





ELAMP Prototype

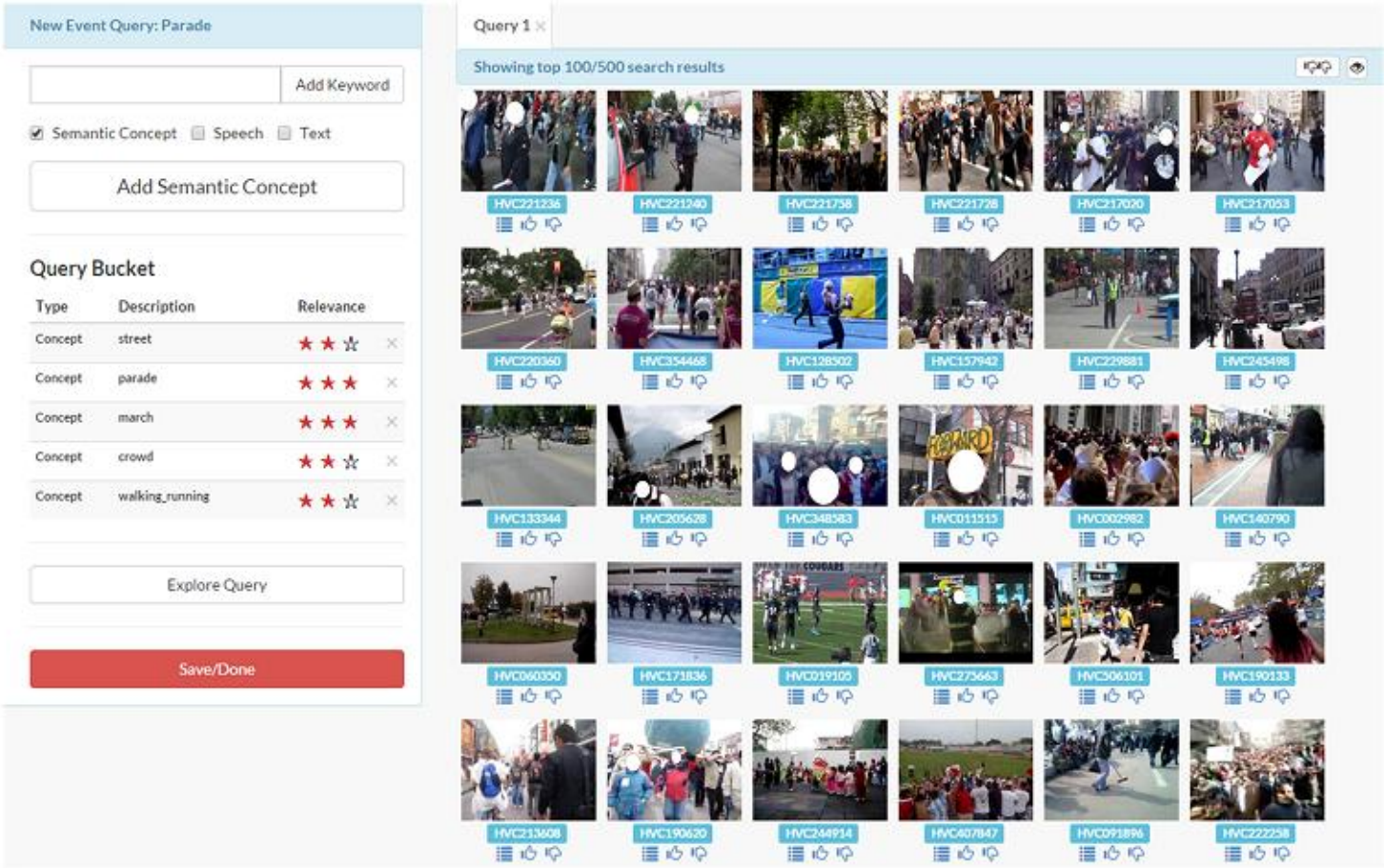


Figure 1: The screenshot of our 0Ex system for the query "E012 Parade". The left panel shows the query bucket that contains relevant visual concepts input by the user. The right panel shows the returned results.

Carnegie Mellon University

Why Important?

Flexible semantic search over the video content:

- Searching personal tour videos:
  - Who is with us in the last year's forest camp?
  - How many beaches have I visited so far? What are they?
- Searching videos in the social streams(Vine/Weibo):
  - What are recent popular videos about the cute dog?
  - What are the user responses to 2015 Super Bowl?
- Video ads:
  - Recommend videos based on the content.
  - Promote videos that have related content (cold start).

Carnegie Mellon University

Why Important?

Flexible semantic search over the video content:

- Searching personal tour videos:
  - Who is with us in the last year's forest camp?
  - How many beaches have I visited so far? What are they?
- Searching videos in the social streams(Vine/Weibo):
  - What are recent popular videos about the cute dog?
  - What are the user responses to 2015 Super Bowl?
- Video ads:
  - Recommend videos based on the content.
  - Promote videos that have related content (cold start).

Carnegie Mellon University

Why Important?

Flexible semantic search over the video content:

- Searching per
  - Who is with
  - How many b
- Searching vid
  - What are rec
  - What are the
- Video ads:
  - Recommend commercial videos based on the content.
  - Promote videos that have related content (cold start).

Carnegie Mellon University

Outline

- Introduction
- **Background**
- Breakthrough 1: Large-scale Semantic Indexing (SIN)
- Breakthrough 2: Self-Paced Reranking (SPaR)
- Conclusions

Representative task: TRECVID Multimedia Event Detection (MED)

Carnegie Mellon University

- Organized and evaluated by the **National Institute of Standards and Technology (NIST).**
- Participants include such as BBN, SRI International, Stanford, Columbia U, UMass, U Maryland, ATT Lab, INRIA, AXES, Fudan U, etc.
- Two benchmarks MED13Test and MED14Test (**32K videos, 20 queries**). Queries are every events such as Birthday party, Changing a vehicle tire, Rock climbing, Making a sandwich, Parade, Parking a vehicle etc.
- Represent the **state-of-the-art performance in both industry and academia**.
- We will discuss the task called **zero-example detection** or **SQ/0Ex**.

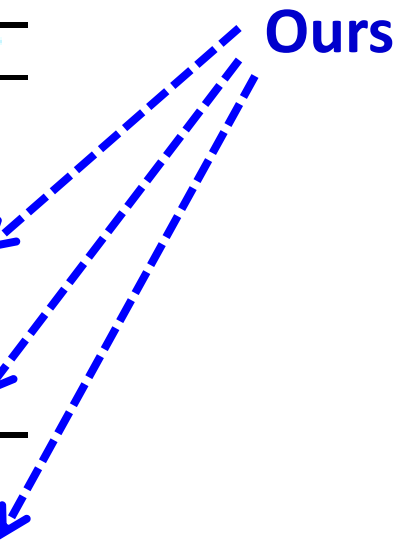
Carnegie Mellon University

Published results on MED13Test

- Published results on the benchmark MED13Test.
- The problem is young and challenging (Mean Average Precision <0.2)

Table 2: MAP (× 100) comparison with the published results on MED13Test.

Method	Year	MAP
SIN/DCNN (visual) [4]	2014	2.5
Composite Concepts [1]	2014	6.4
Tag Propagation [7]	2014	9.6
MMPRF [4]	2014	10.1
Clauses [5]	2014	11.2
Multi-modal Fusion [8]	2014	12.6
SPaR [2]	2014	12.9
Our AutoSQGSys	2015	12.0
Our VisualSys	2015	18.3
Our FullSys	2015	20.7

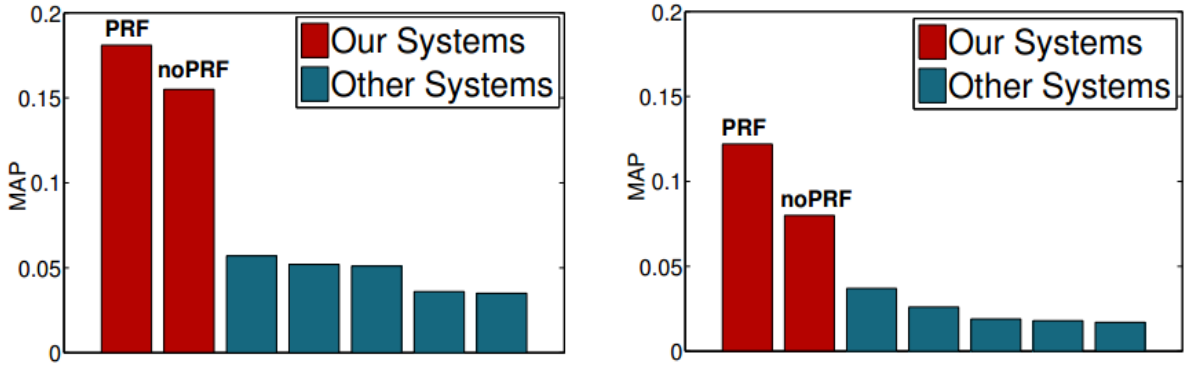


[1] A. Habibian, T. Mensink, and C. G. Snoek. Composite concept discovery for zero-shot video event detection. In ICMR, 2014.
[2] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In MM, 2014.
[4] L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In ICMR, 2014.
[5] H. Lee. Analyzing complex events and human actions in" in-the-wild" videos. In UMD Ph.D Theses and Dissertations, 2014.
[7]M. Mazloom, X. Li, and C. G. Snoek. Few-example video event retrieval using tag propagation. In ICMR, 2014.
[8] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In CVPR, 2014.

Carnegie Mellon University

Results Evaluated by NIST

- Official evaluation by NIST on the largest collection MED14Eval. The evaluation is very rigid because
  - Each system can only submit a single run
  - Within 60 minutes after getting the query
  - The ground-truth data is never released even after the submission
  - Some queries are online generated and unknown to the system beforehand



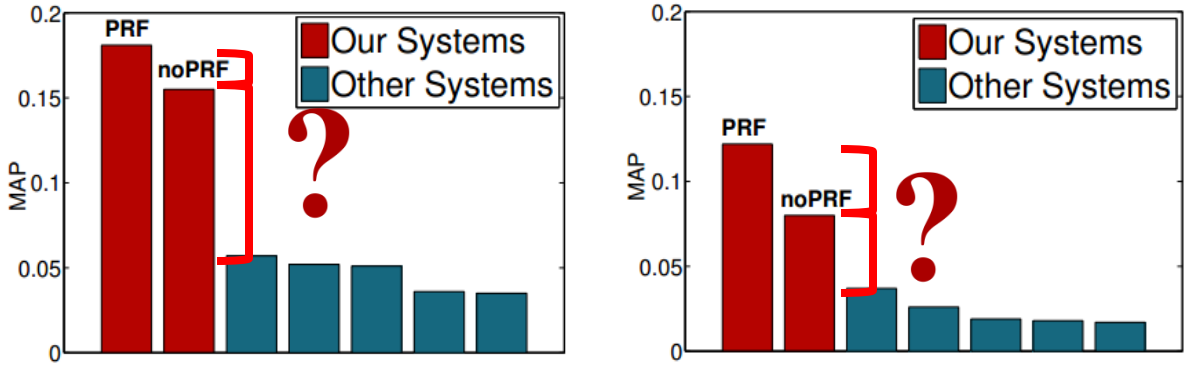
(a) Pre-Specified (PS) (b) Ad-Hoc (AH)

Figure 3: The official results released by NIST TRECVID 2014 on MED14Eval (200, 000 videos).

Carnegie Mellon University

Results Evaluated by NIST

- Official evaluation by NIST on the largest collection MED14Eval. The evaluation is very rigid because
  - Each system can only submit a single run
  - Within 60 minutes after getting the query
  - The ground-truth data is never released even after the submission
  - Some queries are online generated and unknown to the system beforehand

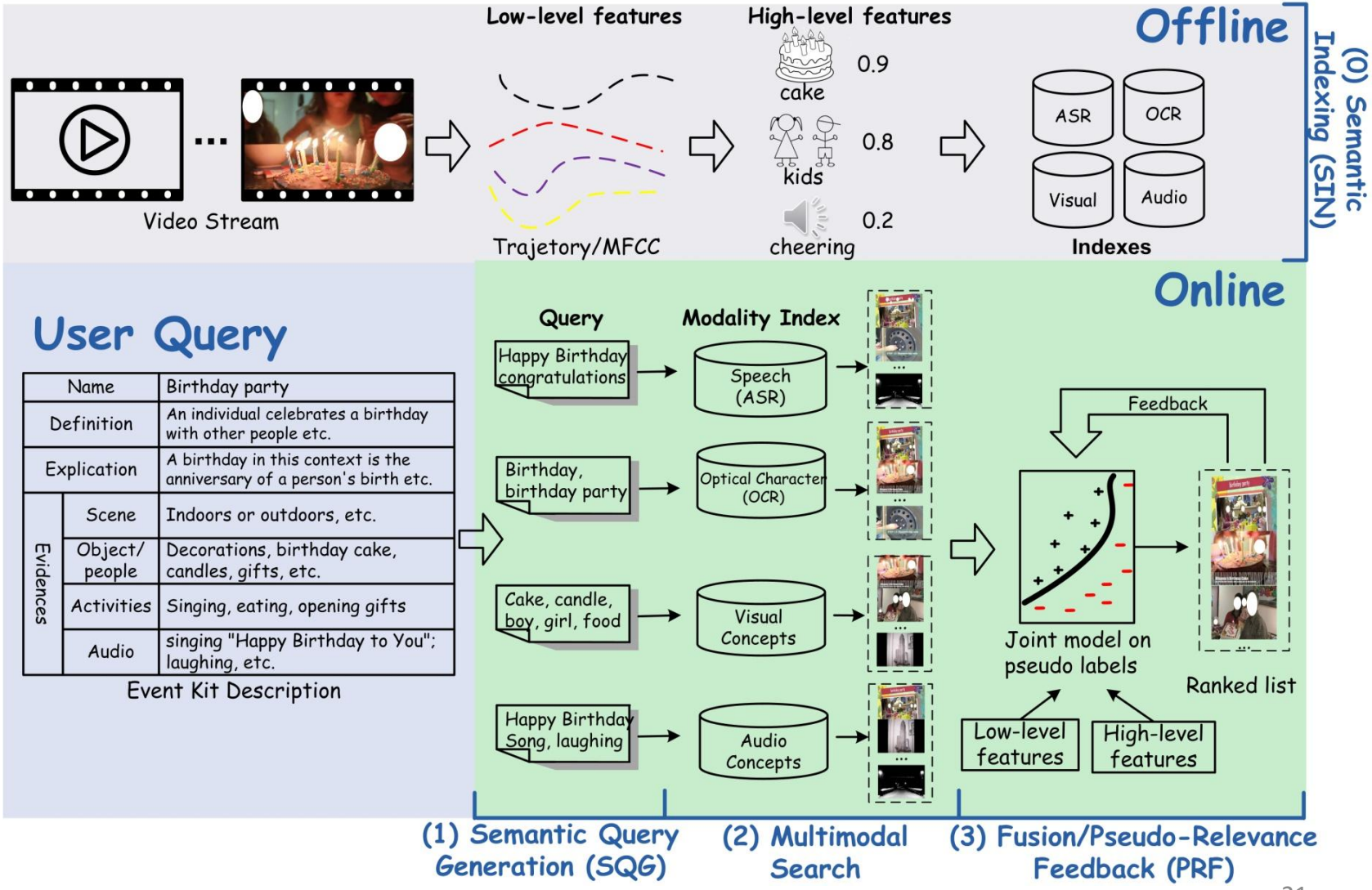


(a) Pre-Specified (PS) (b) Ad-Hoc (AH)

Figure 3: The official results released by NIST TRECVID 2014 on MED14Eval (200, 000 videos).

Carnegie Mellon University

Pipeline



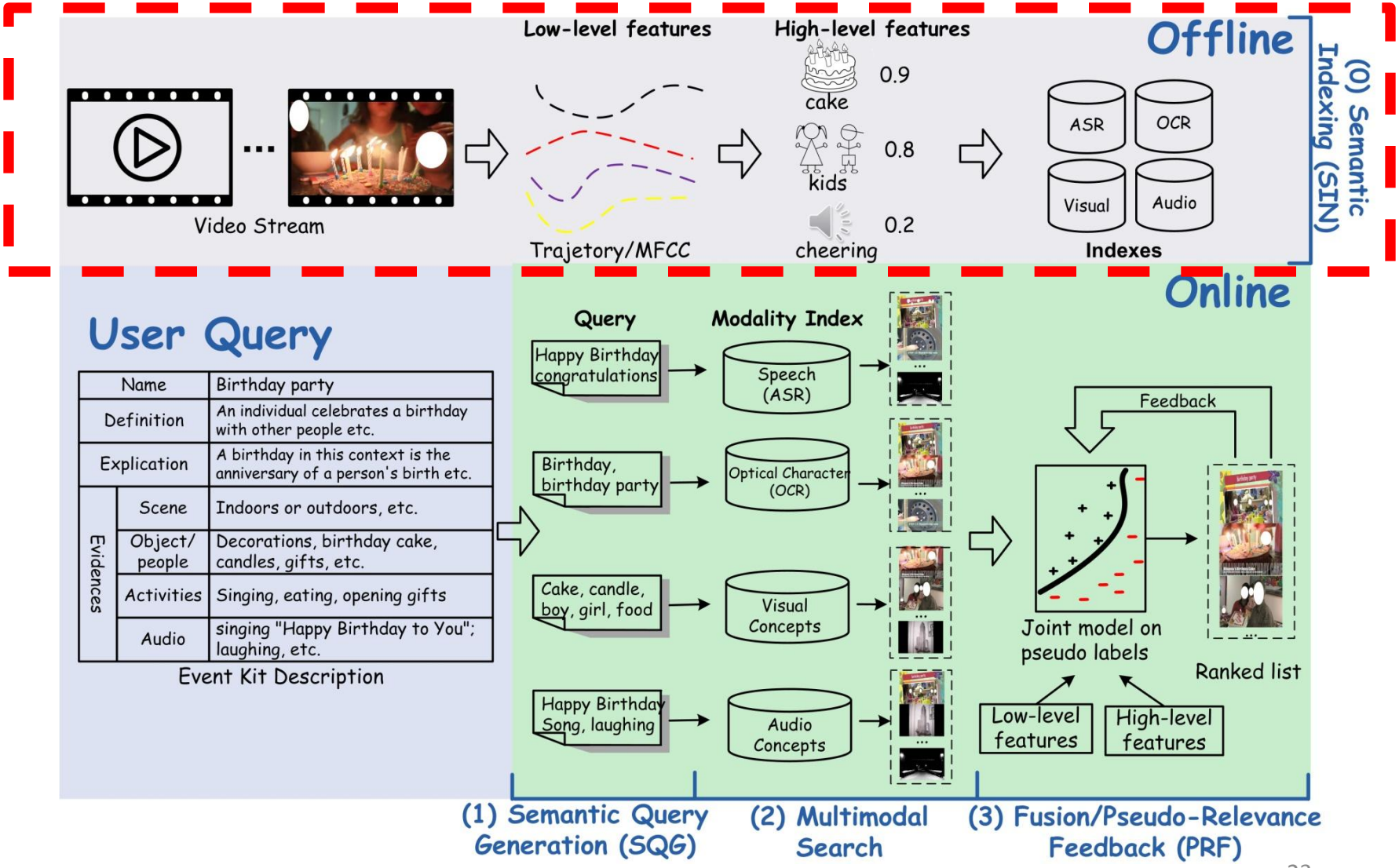
Carnegie Mellon University

Outline

- Introduction
- Background
- **Breakthrough 1: Large-scale Semantic Indexing (SIN)**
- Breakthrough 2: Self-Paced Reranking (SPaR)
- Conclusions

Carnegie Mellon University

Roadmap



Carnegie Mellon University

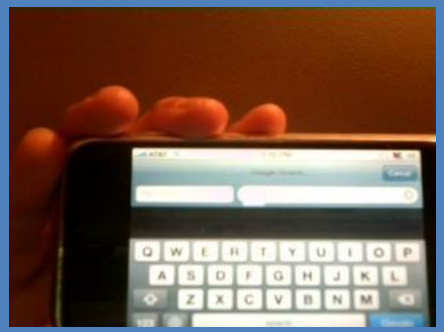
Large-scale Semantic Indexing

- Why do we need shot-based concepts?
  1. **Domain difference in objects.**




In still images






In videos

Carnegie Mellon University

Large-scale Semantic Indexing

- Why do we need shot-based concepts?
  1. Domain difference in objects.
  2. **Actions.**

In still images

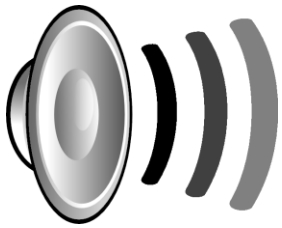
In videos

Carnegie Mellon University

Large-scale Semantic Indexing

- Why do we need shot-based concepts?
  1. Domain difference in objects.
  2. Actions.
  3. **Multimodal detectors (visual & acoustic).**

For the concept diving



water splashing sound

Carnegie Mellon University

Summary of Datasets

Dataset	#Samples	#Classes	Category	Example Concepts
DIY [1]	72,000	1,601	Instructional videos	Yoga, Juggling, Cooking
IACC [2]	600,000	346	Internet archive videos	Baby, Outdoor, Sitting down
YFCC [3]	800,000	609	Amateur videos on Flickr	Beach, Snow, Dancing
Sports [4]	1,100,000	487	Sports videos on YouTube	Bullfighting, Cycling, Skiing

Trained on improved dense trajectory features [5].
Computational challenging: building the semantic detectors costs us more than 1.2 million CPU core hours.

[1] S.-I. Yu, L. Jiang, and A. Hauptmann. Instructional videos for unsupervised harvesting and learning of action examples. In MM, 2014.
[2] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Qu´eenot. TRECVID 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In TRECVID, 2014.
[3] Yahoo! Webscope dataset yfcc-100m. http://labs.yahoo.com/news/yfcc100m/, 2014.
[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, 2014.
[5] H. Wang and C. Schmid. Action recognition with improved trajectories. In ICCV, 2013.

Even more challenging…

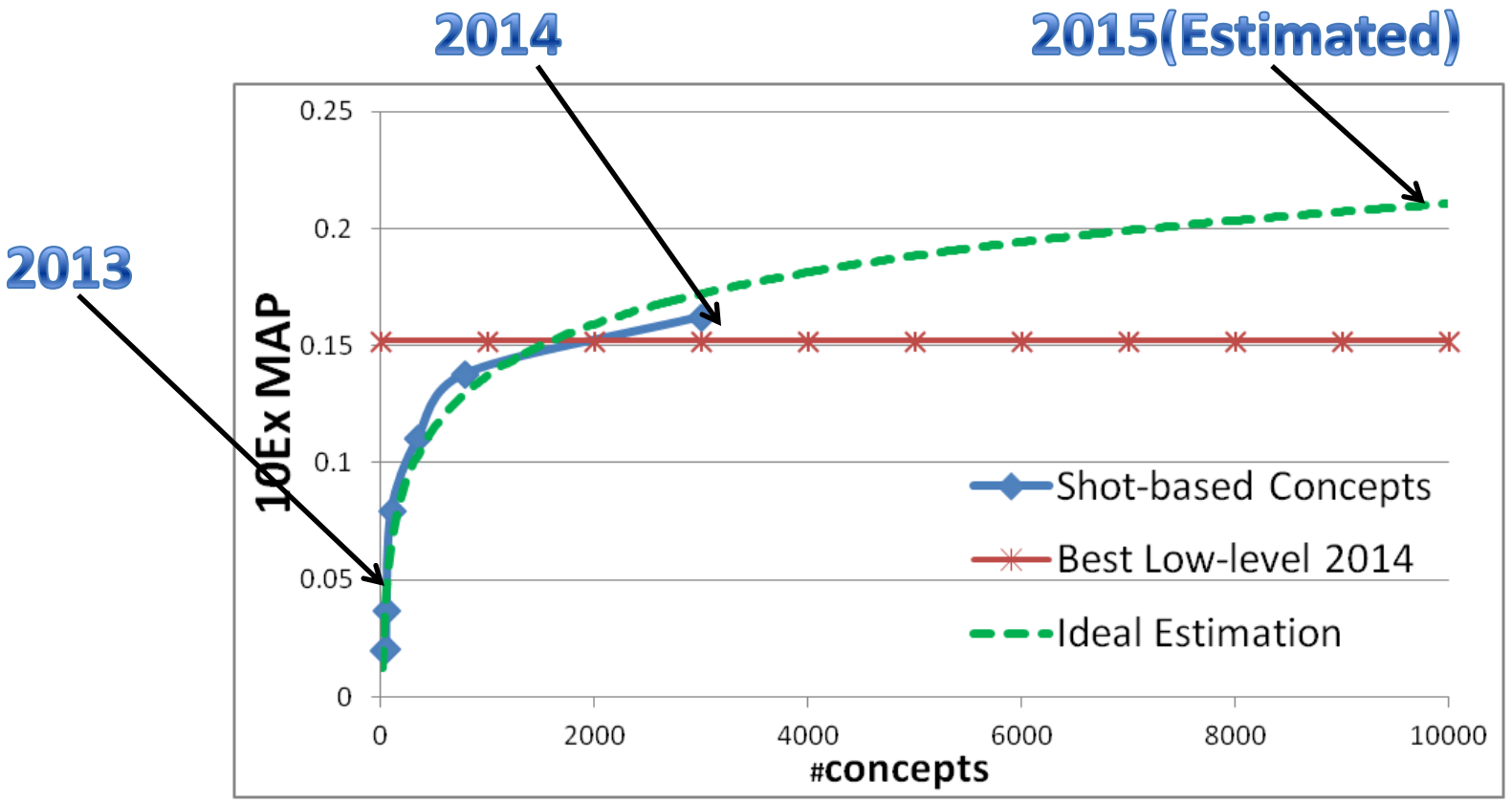
- Noisy (weakly labeled)
- Highly unbalanced data

Our solution: Training detectors iteratively using samples organized in a meaningful sequence (curriculum) [1]

[1] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann. Self-paced curriculum learning. In AAAI, 2015.

Carnegie Mellon University

Preliminary results on MED14Test



The concepts are used as a mid-level representation in 0Ex setting.
Best low-level feature is the improved dense trajectory[].
Shot-based concepts are derived from the weighted fusion.

Carnegie Mellon University

Large-scale Semantic Indexing

- Leveraging multiple modality and PRF.

Table 3: Comparison of modality contribution.

Run	MED13Test		MED14Test	
	1-split	10-splits	1-split	10-splits
FullSys+PRF	22.12	-	22.05	-
FullSys	20.75	19.47±1.19	20.60	18.77±2.16
VisualSys	18.31	18.30±1.11	17.58	17.27±1.82
ASRSys	6.53	6.90±0.74	5.79	4.26±1.19
OCRSys	2.04	7.53±1.01	1.47	4.52±0.91

- Quality and quantity (relevance) of concept detectors which is more important for semantic search?
  - Building **more detectors with reasonable accuracy** seems to be a sensible strategy. Merely increasing the number of low-quality concepts may not improve the performance.

Carnegie Mellon University

Semantic Feature Contributions

Table 4: Comparison of feature contribution.

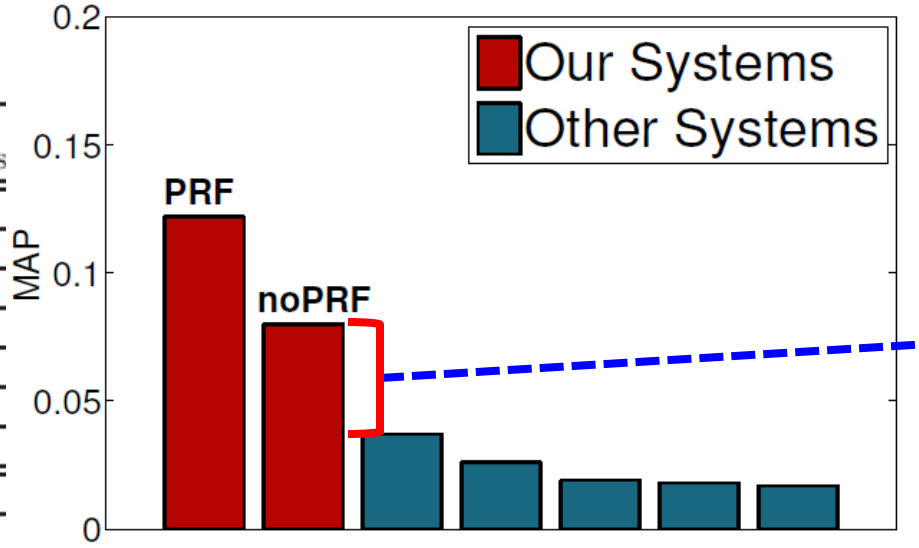
SysID	Visual Concepts					ASR	OCR	MAP		MAP Drop(%)
	sin	sports	yfcc	diy	imagenet			1-split	10-splits	
MED13/IACC		✓	✓	✓	✓	✓	✓	18.93	18.61±1.13	9%
MED13/Sports	✓		✓	✓	✓	✓	✓	15.67	14.68±0.92	25%
MED13/YFCC	✓	✓		✓	✓	✓	✓	18.14	18.47±1.21	13%
MED13/DIY	✓	✓	✓		✓	✓	✓	19.95	18.70±1.19	4%
MED13/ImageNet	✓	✓	✓	✓		✓	✓	18.18	16.58±1.18	12%
MED13/ASR	✓	✓	✓	✓	✓		✓	18.48	18.78±1.10	11%
MED13/OCR	✓	✓	✓	✓	✓	✓		20.59	19.12±1.20	1%
MED14/IACC		✓	✓	✓	✓	✓	✓	18.34	17.79±1.95	11%
MED14/Sports	✓		✓	✓	✓	✓	✓	13.93	12.47±1.93	32%
MED14/YFCC	✓	✓		✓	✓	✓	✓	20.05	18.55±2.13	3%
MED14/DIY	✓	✓	✓		✓	✓	✓	20.40	18.42±2.22	1%
MED14/ImageNet	✓	✓	✓	✓		✓	✓	16.37	15.21±1.91	20%
MED14/ASR	✓	✓	✓	✓	✓		✓	18.36	17.62±1.84	11%
MED14/OCR	✓	✓	✓	✓	✓	✓		20.43	18.86±2.20	1%

Contribution: Sports > ImageNet > ASR > IACC > YFCC > DIY > OCR

Data size: Sports > ImageNet > YFCC > IACC > DIY

Training detectors on "big data" is a promising but challenging solution.

Semantic Feature Contributions



SysID	sin	s						splits	MAP Drop(%)
MED13/IACC								±1.13	9%
MED13/Sports	✓							±0.92	25%
MED13/YFCC	✓							±1.21	13%
MED13/DIY	✓							±1.19	4%
MED13/ImageNet	✓							±1.18	12%
MED13/ASR	✓							±1.10	11%
MED13/OCR	✓							±1.20	1%
MED14/IACC								±1.95	11%
MED14/Sports	✓							±1.93	32%
MED14/YFCC	✓	✓		✓	✓	✓	✓	20.05	18.55±2.13	3%
MED14/DIY	✓	✓	✓		✓	✓	✓	20.40	18.42±2.22	1%
MED14/ImageNet	✓	✓	✓	✓		✓	✓	16.37	15.21±1.91	20%
MED14/ASR	✓	✓	✓	✓	✓		✓	18.36	17.62±1.84	11%
MED14/OCR	✓	✓	✓	✓	✓	✓		20.43	18.86±2.20	1%

Contribution: Sports > ImageNet > ASR > IACC > YFCC > DIY > OCR

Data size: Sports > ImageNet > YFCC > IACC > DIY

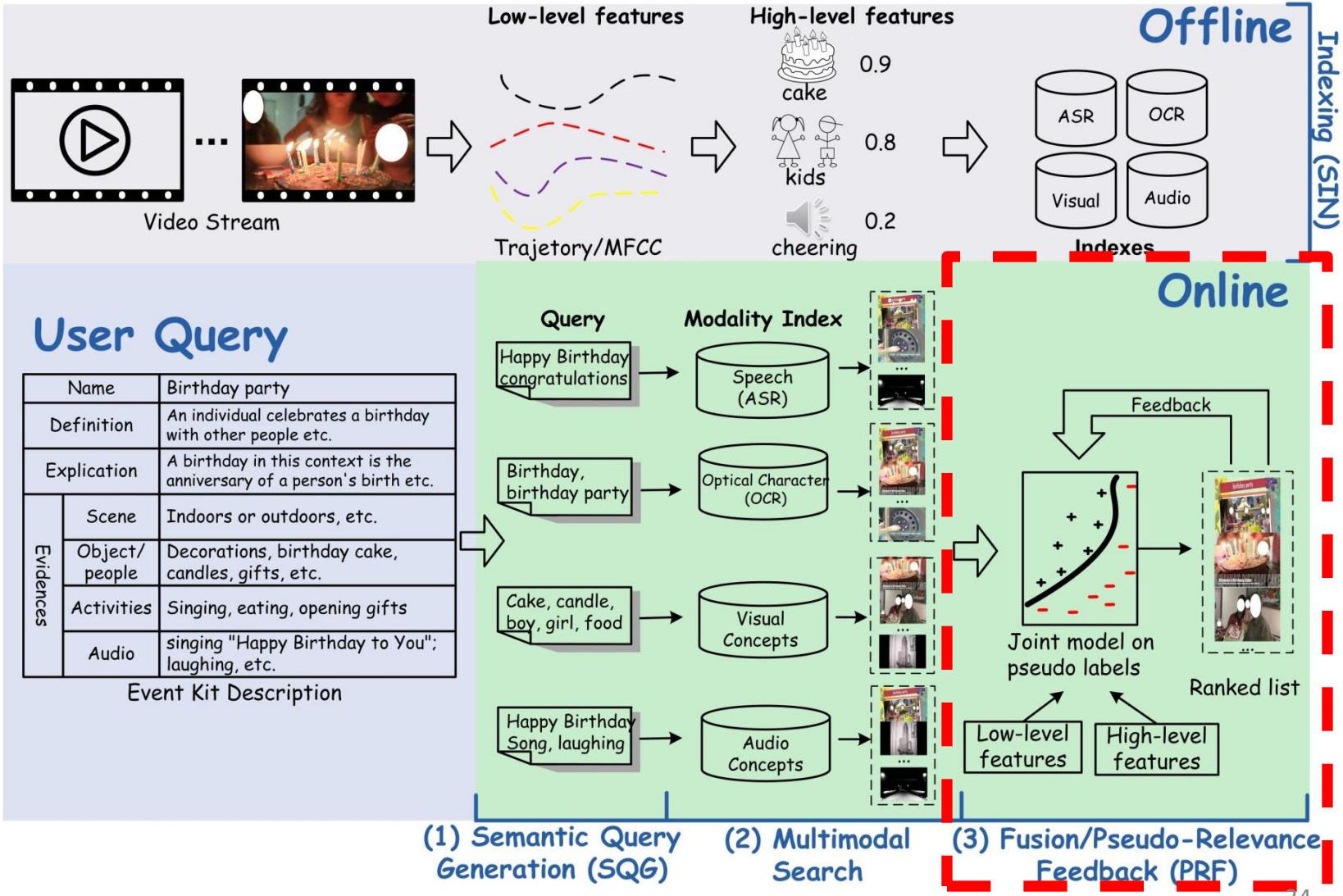
Training detectors on "big data" is a promising but challenging solution.

Carnegie Mellon University

Outline

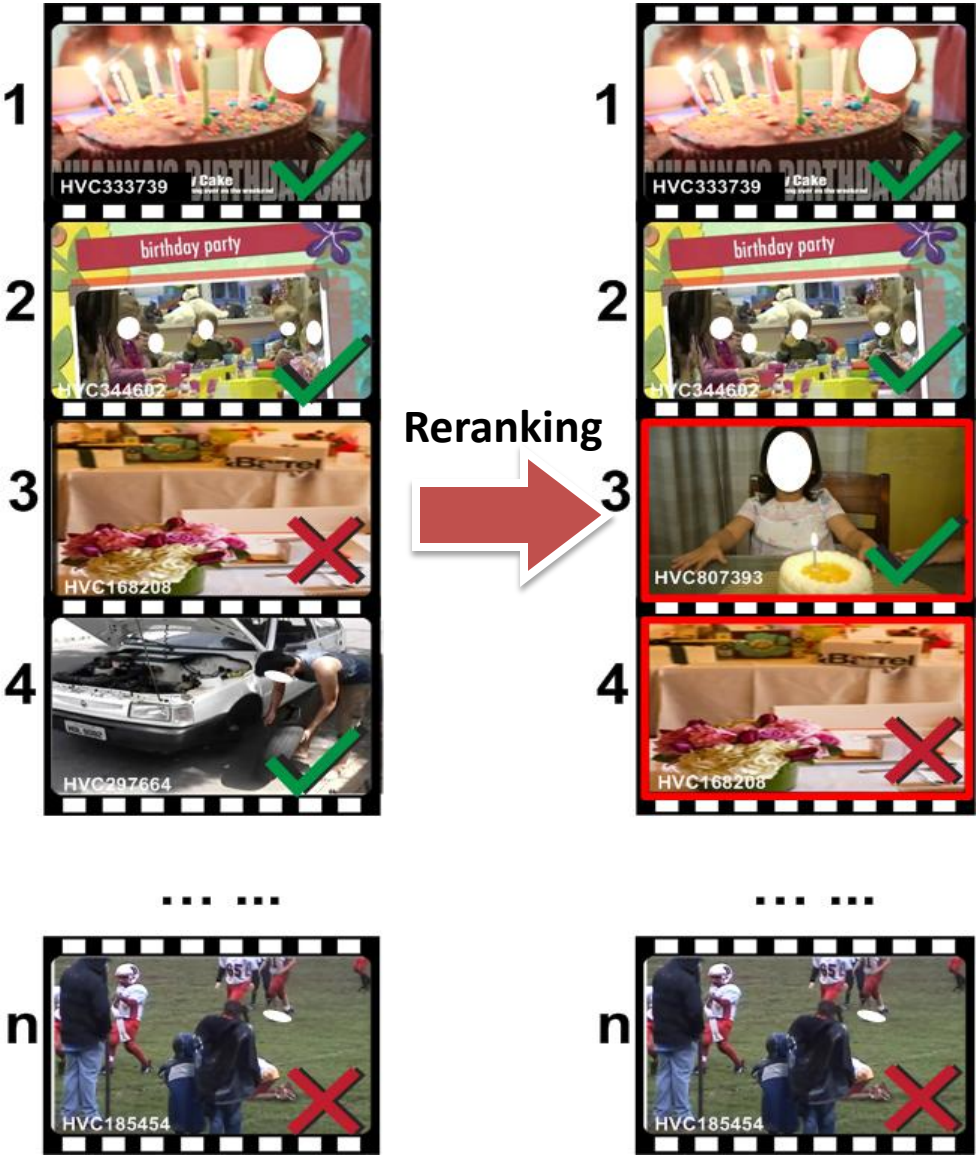
- Introduction
- Background
- Breakthrough 1: Large-scale Semantic Indexing (SIN)
- **Breakthrough 2: Self-Paced Reranking (SPaR)**
- Conclusions

Roadmap


Low-level features
High-level features
Offline
(0) Semantic Indexing (SIN)
Video Stream
Trajetory/MFCC
cake
0.9
kids
0.8
cheering
0.2
ASR
OCR
Visual
Audio
Indexes
User Query
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
Online
Query
Modality Index
Happy Birthday congratulations
Speech (ASR)
Birthday, birthday party
Optical Character (OCR)
Cake, candle, boy, girl, food
Visual Concepts
Happy Birthday Song, laughing
Audio Concepts
Feedback
Joint model on pseudo labels
Ranked list
Low-level features
High-level features
(1) Semantic Query Generation (SQG)
(2) Multimodal Search
(3) Fusion/Pseudo-Relevance Feedback (PRF)

Carnegie Mellon University

Reranking



- Intuition: initial ranked result is noisy.
- Refined by the multimodal info residing in the top ranked videos/images.
- Called **Reranking** or Pseudo Relevance Feedback (PRF).

Carnegie Mellon University

Generic Reranking Algorithm

1: $t = 0$; //Iteration zero
2: Choose the initial pseudo labels and weights;
3: **while** $t \leq$ max iteration **do**
4: Train a reranking model on the fixed labels and weights;
5: Update the pseudo labels and weights;
6: **if** t is small **then** add more pseudo positives;
7: **end while**
8: **return** The list of samples after reranking;

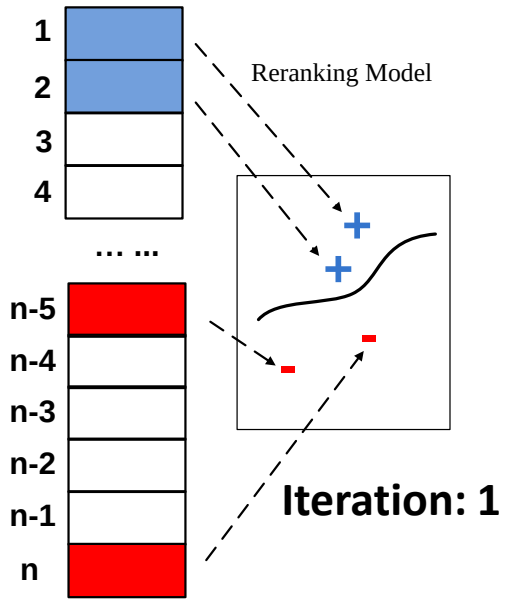
Pseudo labels: assumed (hidden) labels for samples.

Zero-example: ground truth label unknown.

Carnegie Mellon University

Generic Reranking Algorithm

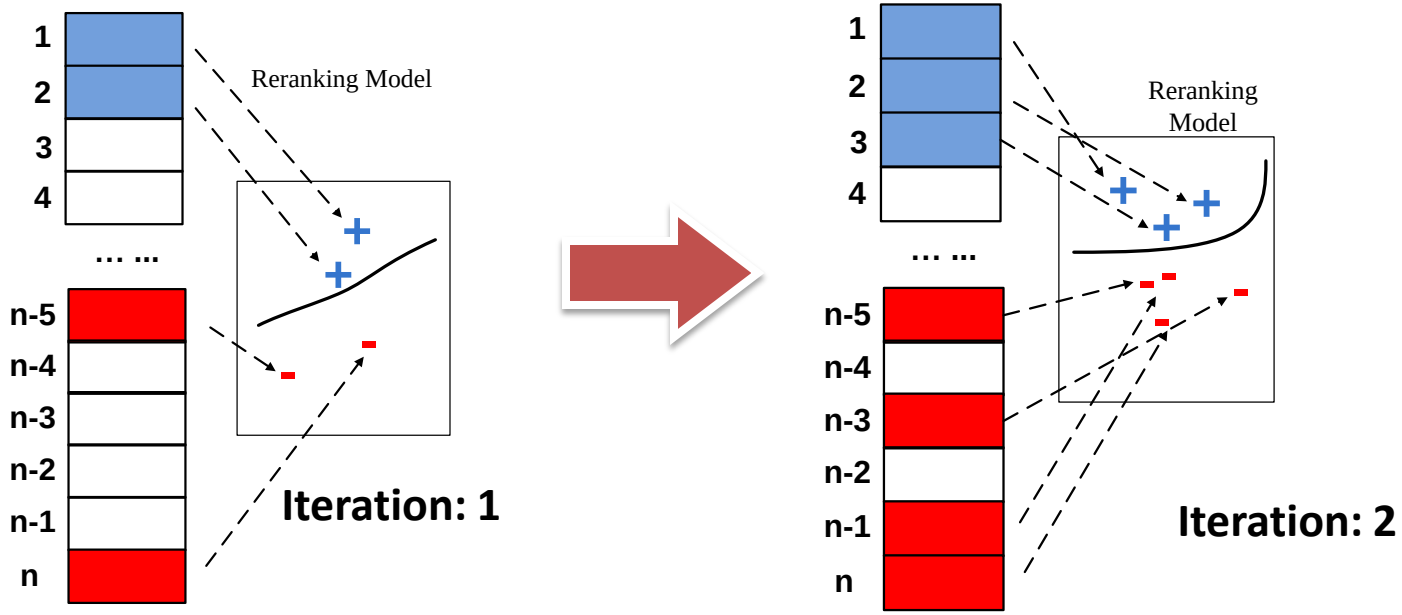
1: $t = 0$; //Iteration zero
2: Choose the initial pseudo labels and weights;
3: **while** $t \leq$ max iteration **do**
4: Train a reranking model on the fixed labels and weights;
5: Update the pseudo labels and weights;
6: **if** t is small **then** add more pseudo positives;
7: **end while**
8: **return** The list of samples after reranking;



Carnegie Mellon University

Generic Reranking Algorithm

1: $t = 0$; //Iteration zero
2: Choose the initial pseudo labels and weights;
3: **while** $t \leq$ max iteration **do**
4: Train a reranking model on the fixed labels and weights;
5: Update the pseudo labels and weights;
6: **if** t is small **then** add more pseudo positives;
7: **end while**
8: **return** The list of samples after reranking;



Carnegie Mellon University

Intuition

True Label	Weighting: Binary	Predefined	Learned
+1	1.0	1.0	1.0
+1	1.0	1/2	1.0
-1	1.0	1/3	0.6
-1	1.0	1/4	0.1

- Existing methods assign equal weights to pseudo samples.
- Intuition: samples ranked at the top are generally more relevant than those ranked lower.
- Our approach: **learn the weight together with the reranking model**.

L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In MM, 2014.

Reranking in Optimization and Conventional Perspective

Carnegie Mellon University

$t = 0$; //Iteration zero
Choose starting values for $\mathbf{y}, \mathbf{v}$;
while $t \leq$ max iteration **do**
  $\Theta_1^{(t+1)}, ..., \Theta_m^{(t+1)} = \arg\max \mathbb{E}_{\mathbf{y},\mathbf{v}}(\Theta_1^{(t)}, ..., \Theta_m^{(t)}; C)$;
  $\mathbf{y}^{(t+1)}, \mathbf{v}^{(t+1)} = \arg\max \mathbb{E}_{\Theta}(\mathbf{y}^{(t)}, \mathbf{v}^{(t)}; k)$;
  if t is small **then** increase $1/k$;
end while
return $[v_1 y_1, \cdots, v_n y_n]^T$;

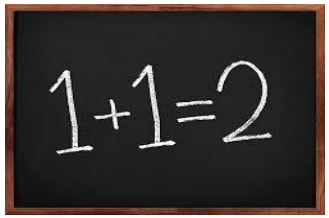
Algorithm 1: Reranking in Optimization Perspective.

CCM (Cyclic Coordinate Method) is used to solve the problem.
Fixing one variable and optimizing the other variables.

Carnegie Mellon University

Self-paced Reranking

- Easy samples to complex samples.
  - Easy sample → smaller loss to the already learned model.
  - Complex sample → bigger loss to the already learned model.



easy as
1 2 3

Age

$$\frac{1}{g-kv}\frac{dv}{dt} = 1$$

$$\int_0^T \frac{1}{g-kv}\frac{dv}{dt}\,dt = \int_0^T dt$$

$$\int_{v_0}^{v(T)} \frac{1}{g-kv}\,dv = T$$

$$-\frac{1}{k}\ln|g-kv|\,\Big|_{v_0}^{v(T)} = T$$

$$\ln\left|\frac{g-kv(T)}{g-kv_0}\right| = -kT$$

$$\frac{g-kv(T)}{g-kv_0} = e^{-kT}$$

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML, 2009.*

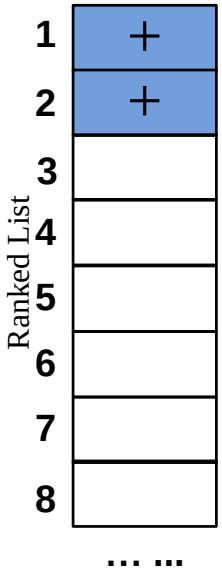
M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In NIPS, 2010.

Carnegie Mellon University

Self-paced Reranking

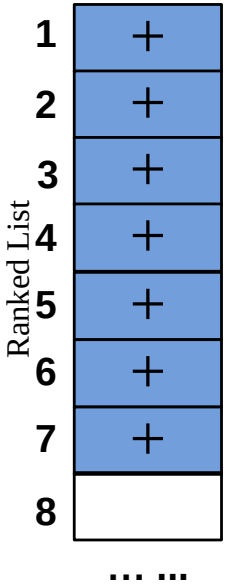
- In the context of reranking : **easy samples are the top-ranked videos that have smaller loss**.

Ranked list of iteration 1



Age

Ranked list of iteration n



Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML, 2009.*

M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In NIPS, 2010.

Carnegie Mellon University

Reranking Results on MED13Test

- Standard setting on the MED13Test set.
- Experiments are conducted on both NIST's split and randomly generated 10 splits.

Table 1: MAP (× 100) comparison with the baseline methods across 20 Pre-Specified events.

Method	NIST's split	10 splits
Without Reranking	3.9	4.9 ± 1.6
Rocchio	5.7	7.4 ± 2.2
Relevance Model	2.6	3.4 ± 1.0
CPRF	6.4	8.3 ± 1.8
Learning to Rank	3.4	4.2 ± 1.4
MMPRF	10.1	13.6 ± 2.4
SPaR	**12.5**	**14.9 ± 2.6**

The results are **statistically significant** at the p-level of 0.05

Reranking Results on MED13Test

- Standard setting on the MED13Test set.
- Experiments are conducted on both NIST's split and randomly generated 10 splits.

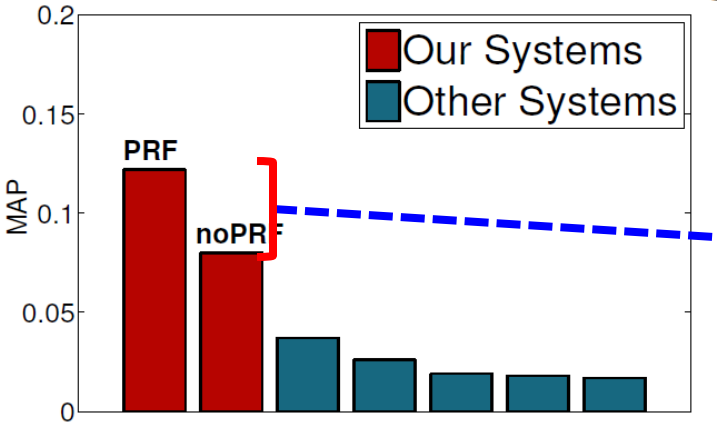


Table 1: MAP (× 100) comparison with the baseline methods across 20 Pre-Specified events.

Method	NIST's split	10 splits
Without Reranking	3.9	4.9 ± 1.6
Rocchio	5.7	7.4 ± 2.2
Relevance Model	2.6	3.4 ± 1.0
CPRF	6.4	8.3 ± 1.8
Learning to Rank	3.4	4.2 ± 1.4
MMPRF	10.1	13.6 ± 2.4
SPaR	**12.5**	**14.9 ± 2.6**

Carnegie Mellon University

Outline

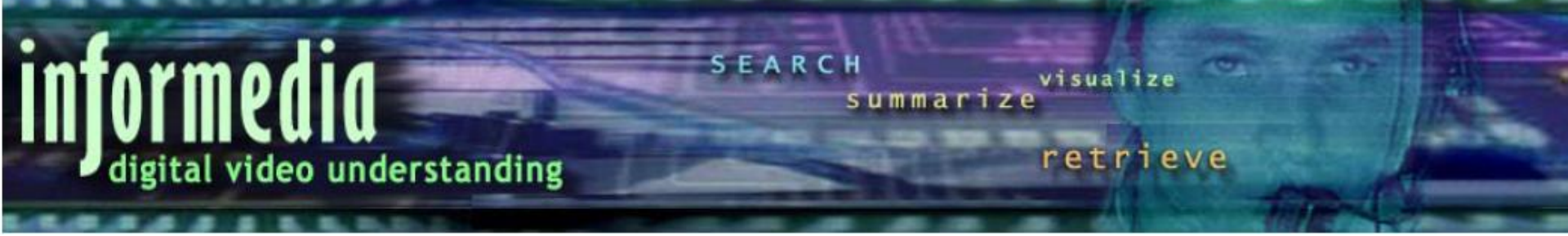
- Introduction
- Background
- Breakthrough 1: Large-scale Semantic Indexing (SIN)
- Breakthrough 2: Self-Paced Reranking (SPaR)
- **Conclusions**

Carnegie Mellon University

Conclusions

A few messages to **take away from this talk**:

- Semantic search in video is a novel, challenging and promising research problem.
- We observed breakthroughs in two directions:
  - Training concept detectors on the large-scale video collection improves the performance. Given limited resources, building more detectors with reasonable accuracy seems to be a sensible strategy. Merely increasing the number of low-quality concepts may not improve the performance.
  - Reranking is a cost-effective approach to improve the search result.
- We plan to share our detectors on the two benchmarks: MED13Test and MED14Test. (send us an email if you are interested in approaching this problem)


informedia
digital video understanding
SEARCH
summarize
visualize
retrieve

THANK YOU.
Q&A?