Web-scale Multimedia Search for Internet Video Content

Lu Jiang
Language Technologies Institute, Carnegie Mellon University

October 15th 2015.

Thesis Committee:
Dr. Alex Hauptmann (Co-chair), Carnegie Mellon University
Dr. Teruko Mitamura (Co-chair), Carnegie Mellon University
Dr. Louis-Philippe Morency, Carnegie Mellon University
Dr. Tat-Seng Chua, National University of Singapore

Outline

- Introduction
- Proposed Approaches:
  - Indexing Semantic Features
  - Semantic Search
  - Video Reranking
  - Building Semantic Concepts
- Conclusions
  - Proposed Work: hybrid search

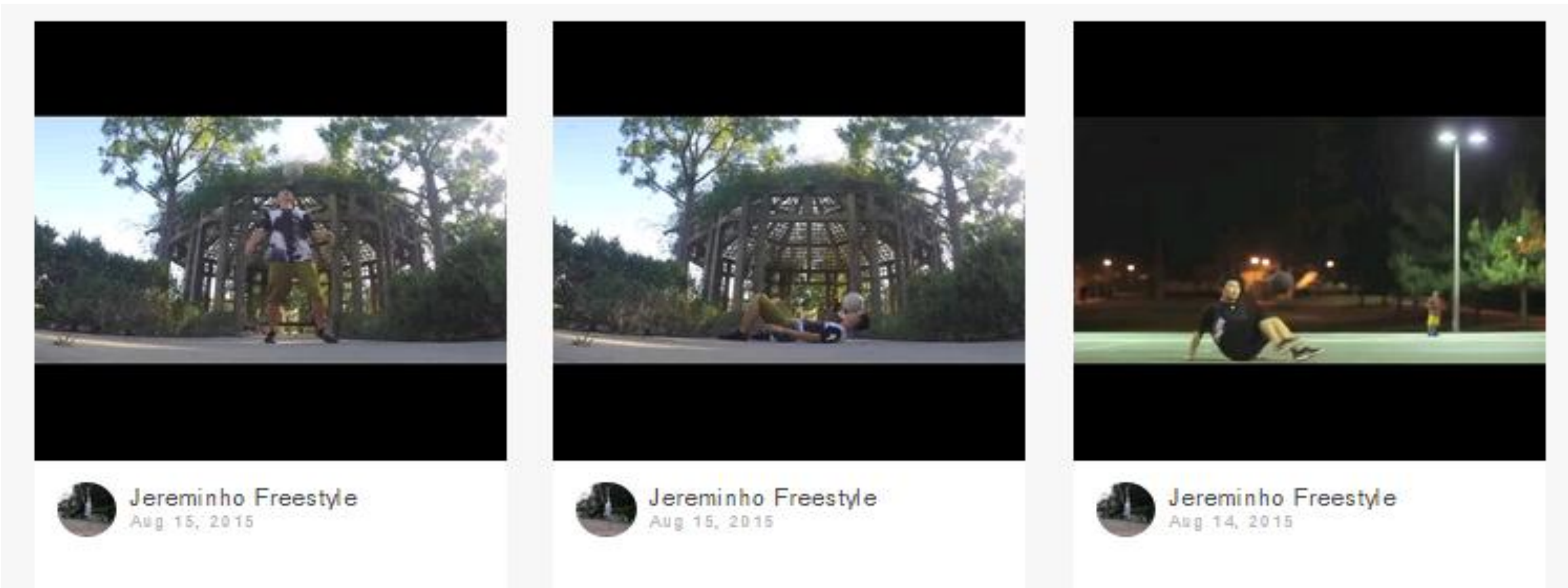
Outline

- **Introduction**
- Proposed Approaches:
  - Indexing Semantic Features
  - Semantic Search
  - Video Reranking
  - Building Semantic Concepts
- Conclusions
  - Proposed Work: hybrid search

Introduction

- We are living in an era of big multimedia data:
  - 300 hours of video are uploaded to YouTube every minute;
  - social media users are posting 12 million videos on Twitter every day;
  - video will account for 80% of all the world's internet traffic by 2019.
- Video search is becoming a valuable source for acquiring information and knowledge.
- Existing large-scale methods are still based on text-to-text matching (user text query to video metadata), which may fail in many scenarios.
  - 66% videos on the social media site Twitter are not associated with hashtag or mention [Vandersmissen et al. 2014]

Introduction



– 66% videos on the social media site Twitter are not associated with hashtag or mention [Vandersmissen et al. 2014]

Introduction

– Much more video captured by mobile phones, surveillance cameras and wearable devices does not have any metadata at all.

Introduction

- We are living in an era of big multimedia data:
  - 300 hours of video are uploaded to YouTube every minute;
  - social media users are posting 12 millions videos on Twitter every day;
  - video will account for 80% of all the world's internet traffic by 2019.
- Video search is becoming a valuable source for acquiring information and knowledge.
- Existing large-scale methods are still based on text-to-

How to acquire information or knowledge in video if there is no way to find it?

may

  - 66% videos on a social media site of Twitter are not associated with meaningful metadata (hashtag or a mention)[Vandersmissen et al. 2014]
  - Much video captured by mobile phones, surveillance cameras and wearable devices does not have any metadata at all.

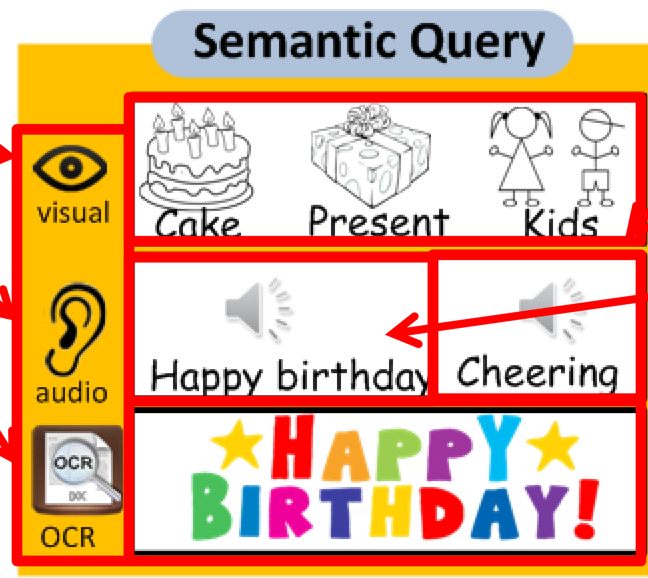
Introduction

- This thesis addresses a fundamental research question: how to satisfy information needs about ***video content*** at a very large scale?
- We embody this question into a concrete content-based video retrieval problem which aims at searching videos solely based on content, without using any user-generated metadata (e.g. titles or descriptions).
- We focus on two types of queries: semantic query and hybrid query.

Semantic Query:

Information need:
Find videos about birthday party.

Multimodal



1) Semantic concept
(acoustic and visual)

2) ASR
(Automatic Speech Recognition)

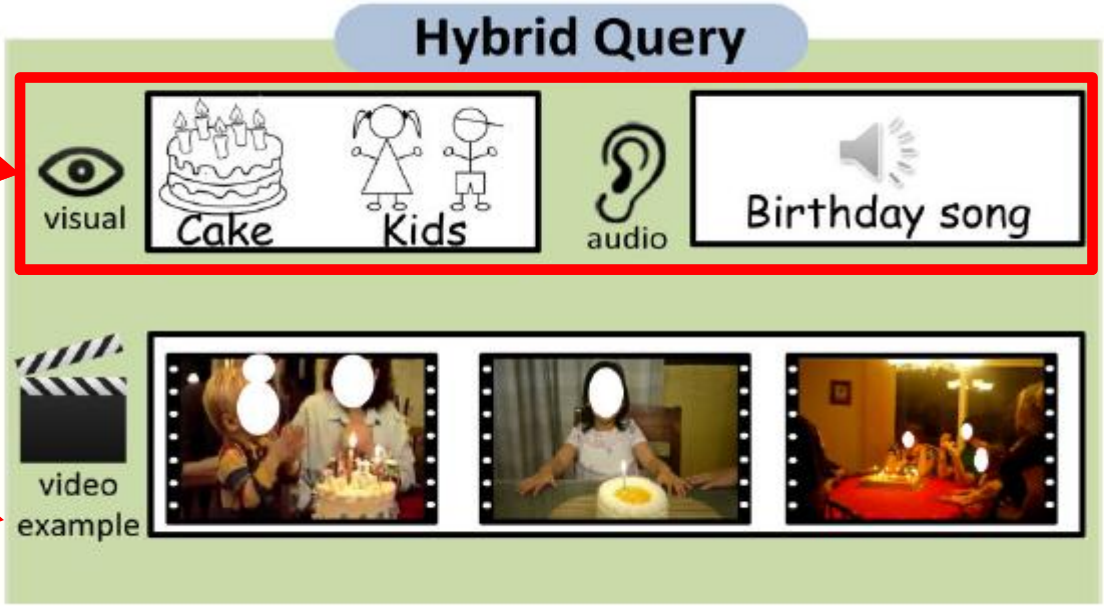
3) OCR
(Optical Character Recognition)

text-to-video search

Hybrid Query:

semantic query

video examples
provided by users



text&video-to-video search

Example Queries

- In response to a, our system should be able to:
  - find simple objects, actions, speech words;
  - search complex activities;

Information need:
people running away after an explosion in urban areas.

Query: **Boolean logical operator**

urban_scene
AND (walking OR running)
OR fire OR smoke
OR audio:explosion
TBefore(audio:explosion, running)

Temporal operators

Example Queries

- Using the query, our system should be able to
  - find simple objects, actions, speech words;
  - search complex activities;
  - answer questions by/in videos.

Information need:
What are they doing?

Query:

person AND action
AND

How to learn Tai Chi Chuan

Challenges

- The problem was initiated by a TRECVID task Multimedia Event Detection (MED) in 2012 (common evaluation benchmark).
  – State-of-the-art accuracy is very low.
  – Large-scale system can only handle 200k videos (5 min to search).
- For this understudied problem, this thesis confronts the following research challenges:
  – Algorithms to boost state-of-the-art accuracy.
  – Efficient methods to search billions of videos.

Preliminary Results

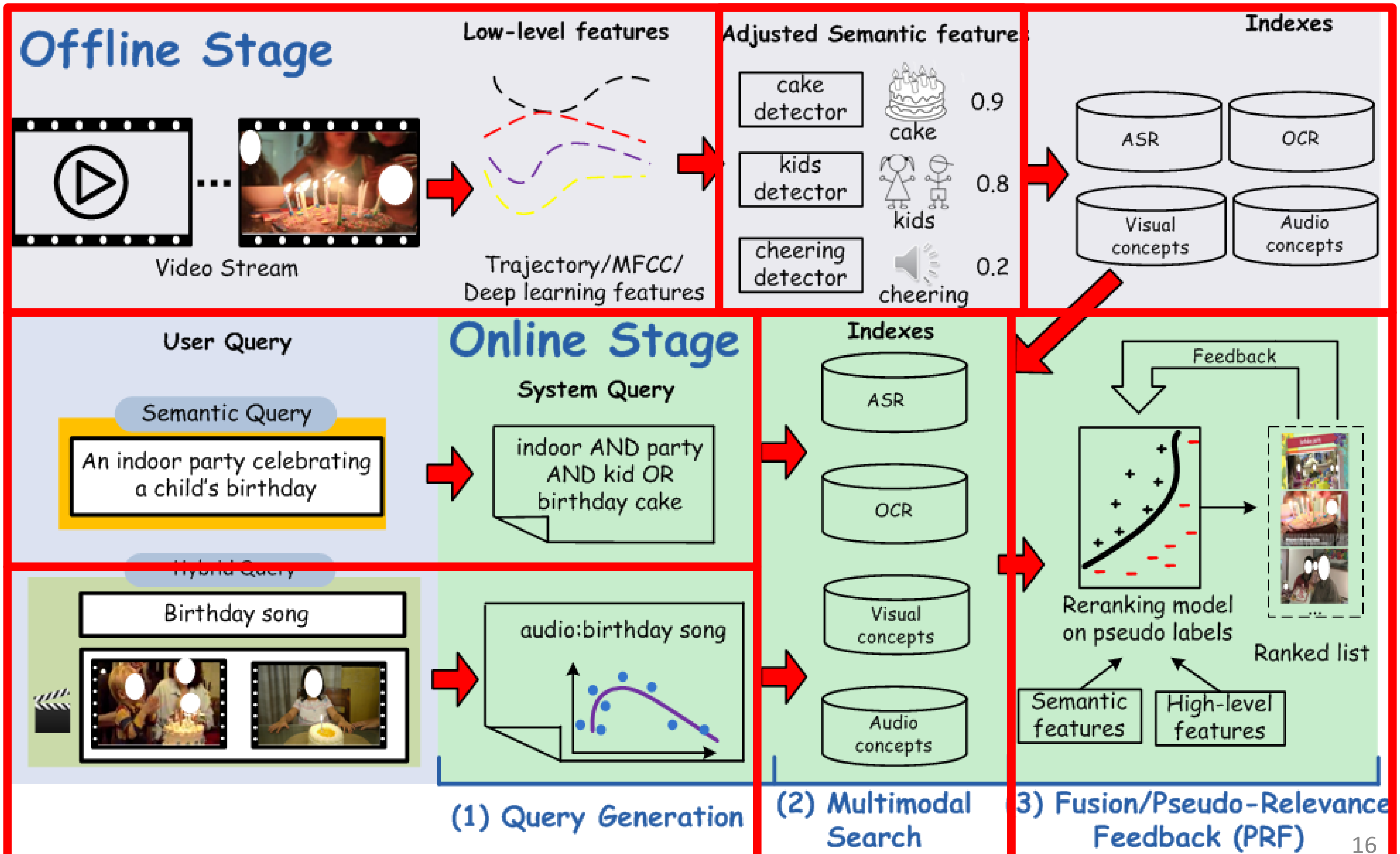
- We proposed a novel and practical solution that can
  - substantially boost state-of-the-art accuracy across a number of datasets.
  - Scale up the search to hundreds of millions of Internet videos.
    - 0.2 second to process a semantic query on 100 million videos
    - 1 second to process a hybrid query on 1 million videos.
- Within a system called E-Lamp Lite, we implemented the first of its kind large-scale multimedia search engine for Internet videos:
  - Achieved the **best accuracy** in TRECVID MED zero-example search 2013, 2014 and 2015, the most representative task on this task. **3x better than** the runner-up in 2014.
  - To the best of our knowledge, it is **the first content-based retrieval system** that can search a collection of 100 million videos.

From large-scale to web-scale

Let the above videos represent the upper-bound of the current largest dataset for this problem (200K videos)

(From Large-scale to Web-scale)

Framework



Offline Stage
Low-level features
Video Stream
Trajectory/MFCC/
Deep learning features
Adjusted Semantic feature
cake detector
cake
0.9
kids detector
kids
0.8
cheering detector
cheering
0.2
Indexes
ASR
OCR
Visual concepts
Audio concepts
User Query
Semantic Query
An indoor party celebrating a child's birthday
Online Stage
System Query
indoor AND party AND kid OR birthday cake
Birthday song
audio:birthday song
Indexes
ASR
OCR
Visual concepts
Audio concepts
Feedback
Reranking model on pseudo labels
Ranked list
Semantic features
High-level features
(1) Query Generation
(2) Multimodal Search
(3) Fusion/Pseudo-Relevance Feedback (PRF)

Key Contributions: First-of-its-kind Framework

- The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR'15]. The proposed framework supports text-to-video, video-to-video, and text&video-to-video search [MM'12]. **(Chapter 1 and 5)**

[ICMR15] Lu Jiang, Shoou-I Yu, Deyu Meng, Teruko Mitamura, Alexander Hauptmann. Bridging the Ultimate Semantic Gap: A Semantic Search Engine for Internet Videos. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2015.
[MM12] Lu Jiang, Alexander Hauptmann, Guang Xiang. Leveraging High-level and Low-level Features for Multimedia Event Detection. *In ACM Multimedia (MM)*, 2012.

Key Contributions: Self-paced curriculums learning theory

- The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR’15]. The proposed framework supports text-to-video, video-to-video, and text&video-to-video search [MM’12].
- A novel theory about self-paced curriculums learning and its application on robust concept detector training [NIPS’14, AAAI’15]. **(Chapter7)**

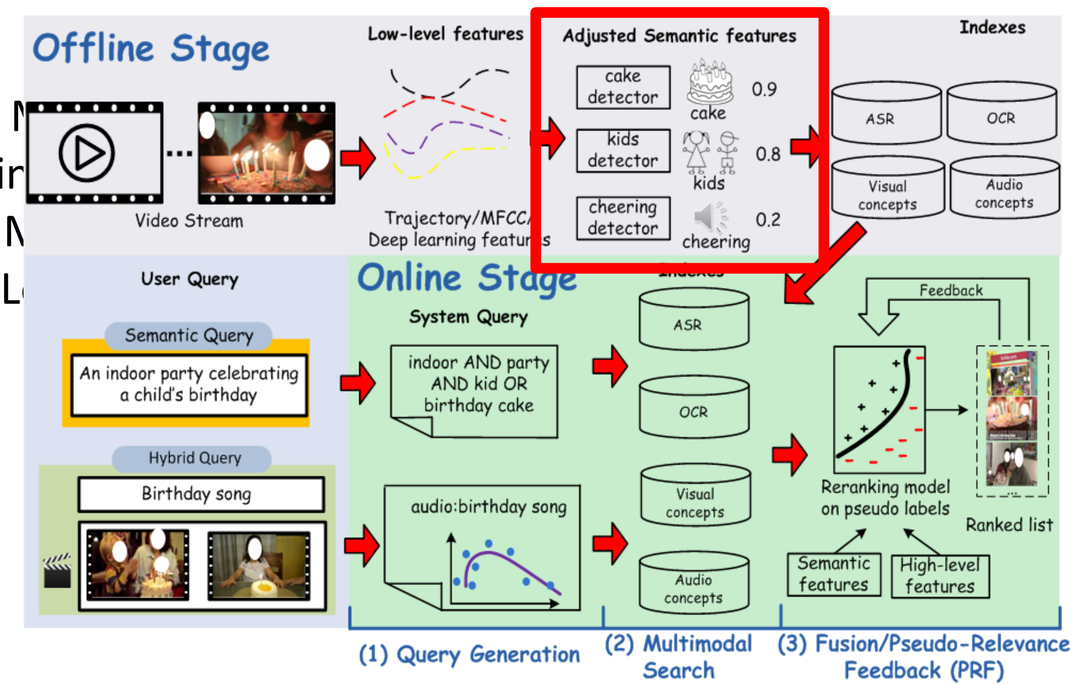
[AAAI15] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, Alexander Hauptmann. Self-paced Curriculum Learning. *In Conference on Artificial Intelligence (AAAI)*, 2015.
[NIPS14] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhen-Zhong Lan, Shiguang Shan, Alexander Hauptmann. Self-paced Learning with Diversity. *In Neural Information Processing Systems (NIPS)*, 2014.

Key Contributions: Self-paced curriculums learning theory

- The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR'15]. The proposed framework supports text-to-video, video-to-video, and text&video-to-video search [MM'12].
- A novel theory about self-paced curriculums learning and its application on robust concept detector training [NIPS'14, AAAI'15]. **(Chapter7)**

[AAAI15] Lu Jiang, Deyu M… …n. Self-paced Curriculum Learnin…

[NIPS14] Lu Jiang, Deyu M… …xander Hauptmann. Self-paced L… …g Systems *(NIPS)*, 2014.

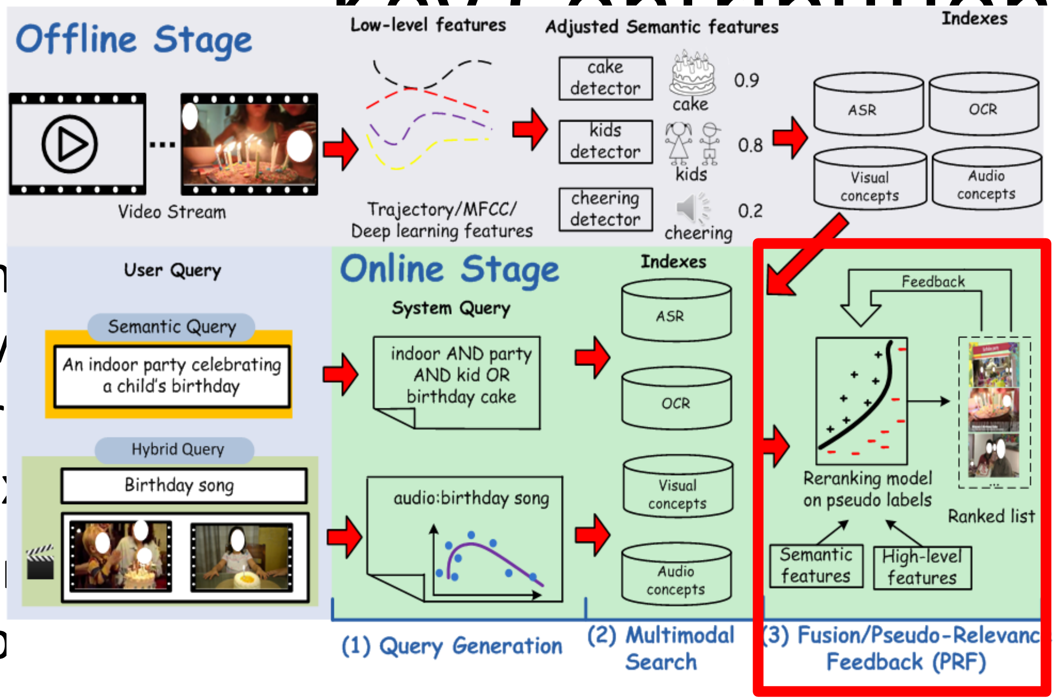


Key Contributions: Reranking

- The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR'15]. The proposed framework supports text-to-video, video-to-video, and text&video-to-video search [MM'12].
- A novel theory about self-paced curriculums learning and its application on robust concept detector training [NIPS'14, AAAI'15].
- Novel reranking algorithms for improving performance. They have concise mathematical objectives to optimize and useful properties that can be theoretically verified [MM'14, ICMR'14]. **(Chapter6)**

[MM14] Lu Jiang, Deyu Meng, Teruko Mitamura, Alexander Hauptmann. Easy Samples First: Selfpaced Reranking for Zero-Example Multimedia Search. *In ACM Multimedia (MM)*, 2014.
[ICMR14] Lu Jiang, Teruko Mitamura, Shoou-I Yu, Alexander Hauptmann. Zero-Example Event Search using MultiModal Pseudo Relevance Feedback. *In ACM International Conference on Multimedia Retrieval (ICMR)*, 2014.

Key Contributions:



- Th ... ontent-based search ov ... MR'15]. The pr ... deo-to-video, and te...
- A ... arning and its ap ... ; [NIPS'14, AAAI'15].
- Novel reranking algorithms for improving performance. They have concise mathematical objectives to optimize and useful properties that can be theoretically verified [MM'14, ICMR'14]. **(Chapter6)**

[MM14] Lu Jiang, Deyu Meng, Teruko Mitamura, Alexander Hauptmann. Easy Samples First: Selfpaced Reranking for Zero-Example Multimedia Search. *In ACM Multimedia (MM)*, 2014.
[ICMR14] Lu Jiang, Teruko Mitamura, Shoou-I Yu, Alexander Hauptmann. Zero-Example Event Search using MultiModal Pseudo Relevance Feedback. *In ACM International Conference on Multimedia Retrieval (ICMR)*, 2014.

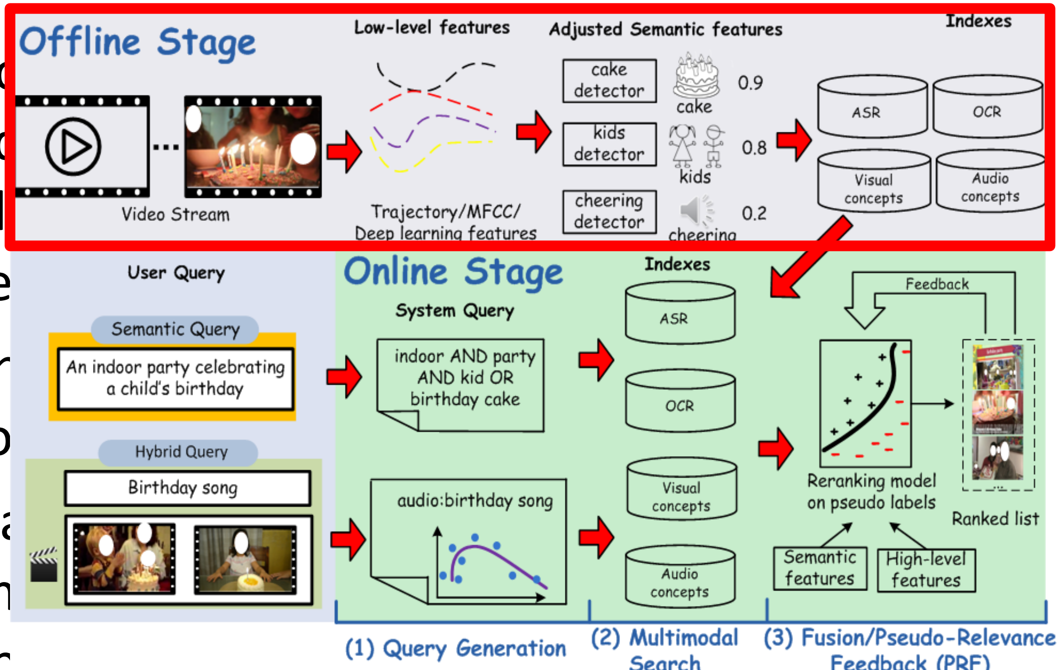
Key Contributions: Scalable Indexing Method

- The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR'15]. The proposed framework supports text-to-video, video-to-video, and text&video-to-video search [MM'12].
- A novel theory about self-paced curriculums learning and its application on robust concept detector training [NIPS'14, AAAI'15].
- Novel reranking algorithms for improving performance. They have concise mathematical objectives to optimize and useful properties that can be theoretically verified [MM'14, ICMR'14].
- A concept adjustment method representing a video by a few salient and consistent concepts that can be efficiently indexed by the modified inverted index [MM'15] **(Chapter3)**

[MM15] Lu Jiang, Shoou-I Yu, Deyu Meng, Yi Yang, Teruko Mitamura, Alexander Hauptmann. Fast and Accurate Content-based Semantic Search in 100M Internet Videos. *In ACM Multimedia (MM)*, 2015

Key Contributions: Scalable Indexing Method

- The first-c … based search over hund …]. The proposed … video, and text&vide …
- A novel th … and its applicatio … 14, AAAI'15].
- Novel rera … . They have concise m … l properties that can b …].



- A concept adjustment method representing a video by a few salient and consistent concepts that can be efficiently indexed by the modified inverted index [MM'15] **(Chapter3)**

[MM15] Lu Jiang, Shoou-I Yu, Deyu Meng, Yi Yang, Teruko Mitamura, Alexander Hauptmann. Fast and Accurate Content-based Semantic Search in 100M Internet Videos. *In ACM Multimedia (MM)*, 2015

Thesis Statement

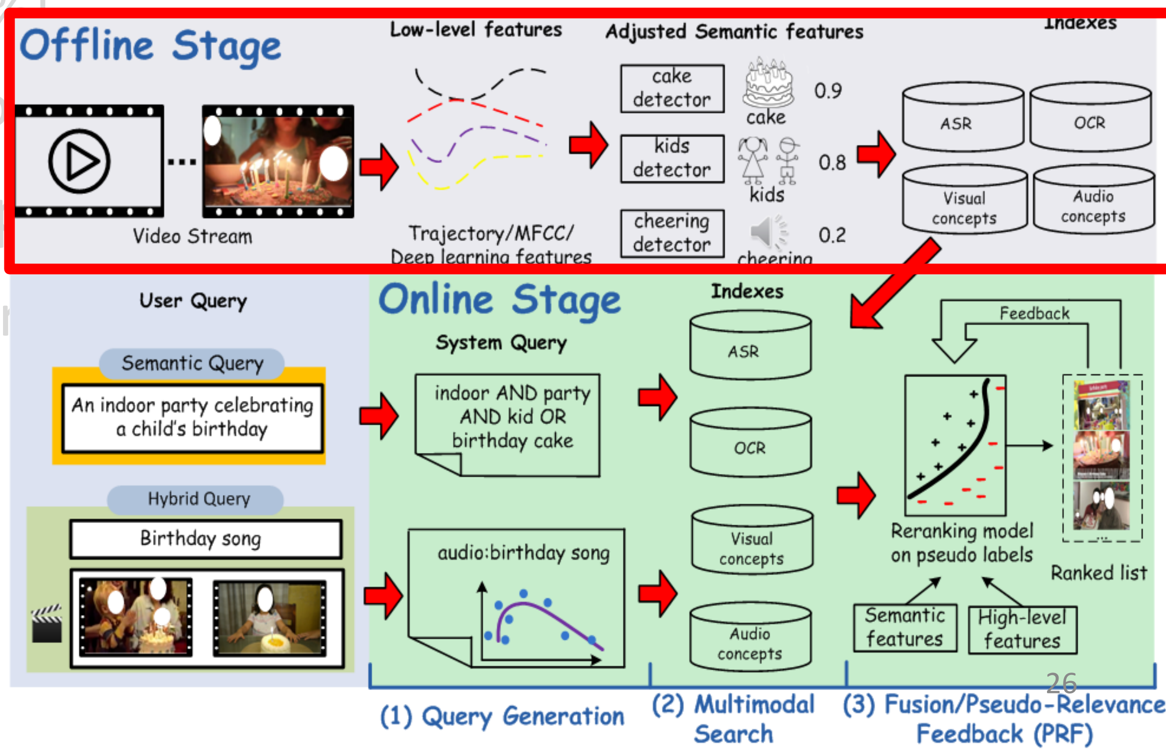
- In this thesis, we approach a fundamental problem of searching information in video content at a very large scale. We address the problem by proposing an accurate, efficient, and scalable method that can search the content of billions of videos by semantic visual/acoustic concepts, speech, visible texts, video examples, or any combination of these elements.

Outline

- Introduction
- **Proposed Approaches:**
  - Indexing Semantic Features [95%]
  - Semantic Search [95%]
  - Video Reranking [95%]
  - Building Semantic Concepts [80%]
- Conclusions
  - Proposed Work: hybrid search [10%]

Outline

- Introduction
- Approaches:
  - **Indexing Semantic Features [95%]**
  - Semantic Search [95%]
  - Video Reranking [95%]
  - Building Semantic Co
- Conclusions and Pro
  - Proposed Work: Hybr [10%]



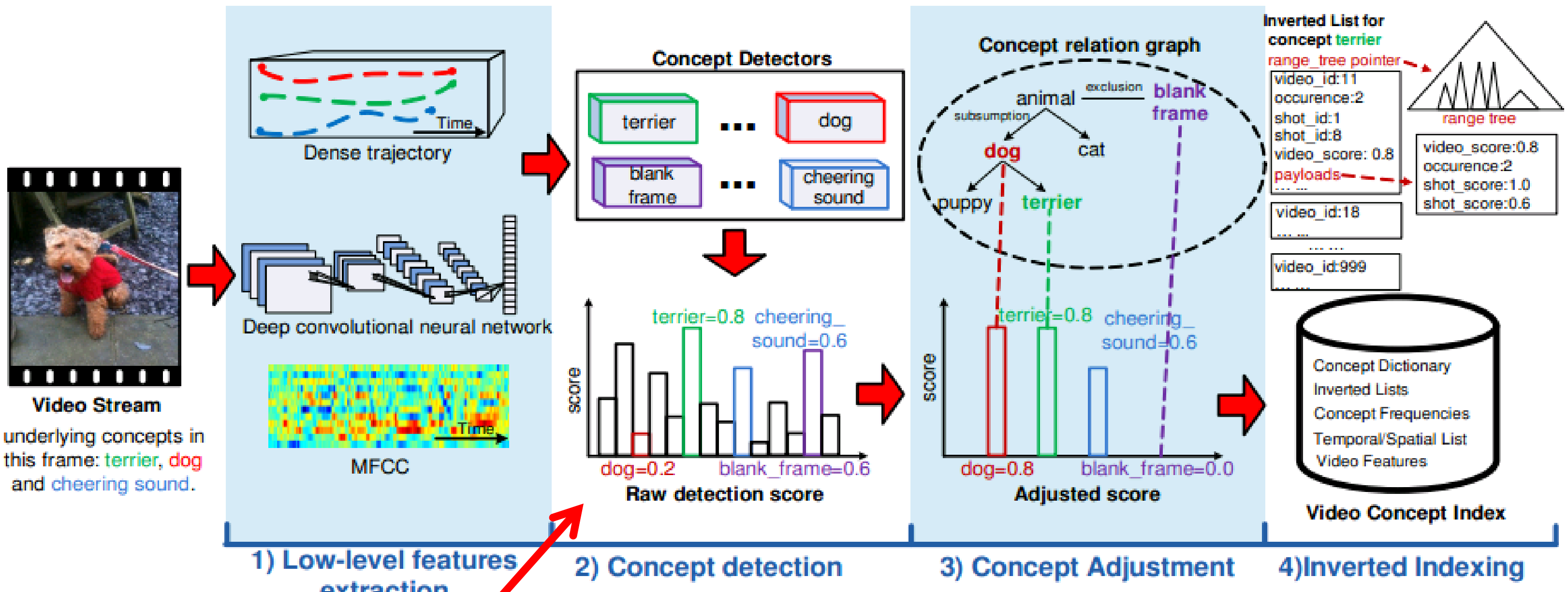
Introduction to Indexing Semantic Features

- Semantic features include ASR (speech), OCR (visible text), visual concepts and audio concepts.
- Indexing textual features like ASR and OCR is well studied.

- Indexing semantic concepts is well studied.
- Existing methods index the raw detection score of semantic concepts by dense matrices [Mazloom et al. 2014][Wu et al. 2014][Lee et al. 2014]
- We propose a scalable semantic concept indexing method. The key is a novel method called concept adjustment.

Masoud Mazloom, Xirong Li, and Cees GM Snoek. Few-example video event retrieval using tag propagation. In *ICMR, 2014.*
Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR, 2014.*
Hyungtae Lee. Analyzing complex events and human actions in" in-the-wild" videos. In *UMD Ph.D Theses and Dissertations, 2014.*

Method Overview

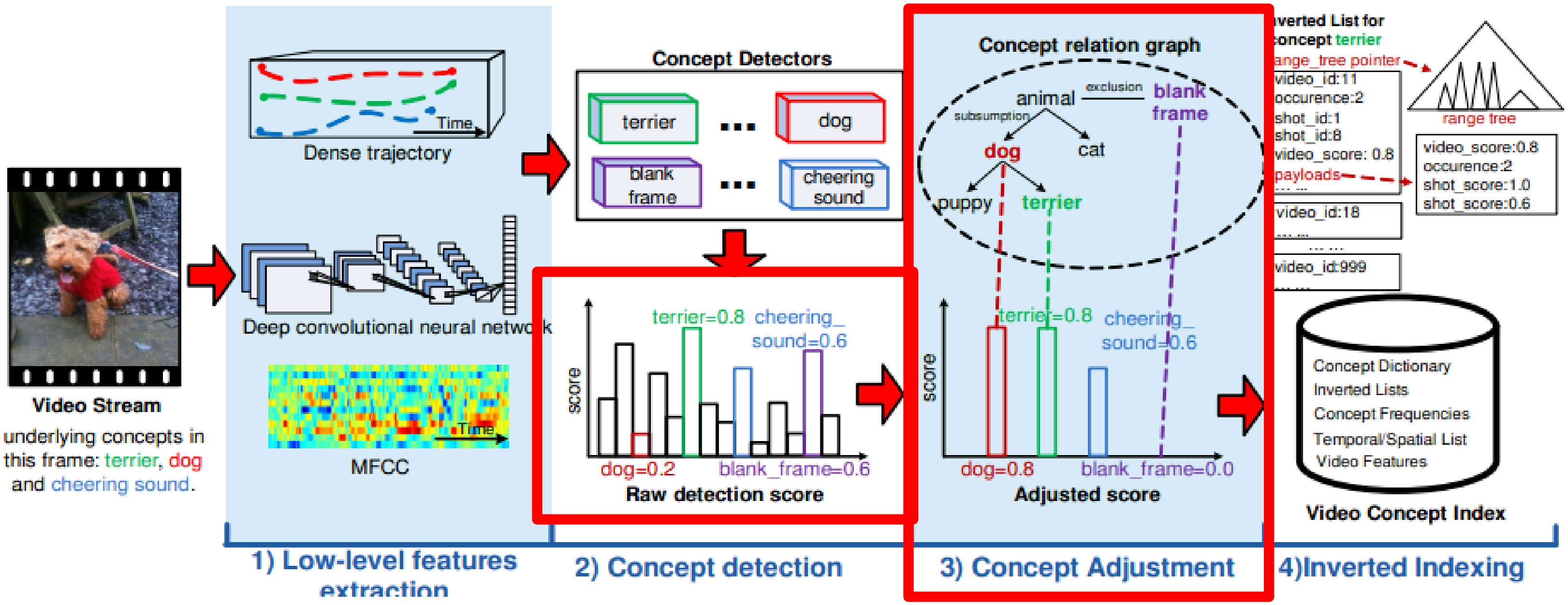


steps in the offline video indexing.

raw score representation

- Represent raw video (or video clip) by low-level features.
- Semantic concept detectors are of limited accuracy. The raw detections are meaningful but very noisy.

Method Overview



- The raw score representation has two problems:
  - **Distributional inconsistency**: every video has every concept in the vocabulary (with a small but nonzero score);
  - **Logical inconsistency**: a video may contain a "terrier" but not a "dog".
- To address the problems, we introduce a novel step called concept adjustment which represents a video by **a few salient and logically consistent visual/audio concepts**.

Concept Adjustment Model

- The proposed adjustment model is:

distributional consistency

$$\arg\min_{\mathbf{v}\in[0,1]^m} \frac{1}{2}\|\mathbf{v} - f_p(\mathbf{D})\|_2^2 + \boxed{g(\mathbf{v};\alpha,\beta)}$$
$$\text{subject to } \boxed{\mathbf{A}\mathbf{v} \leq \mathbf{c}}$$

logical consistency

where $\mathbf{v} \in \mathbb{R}^{m\times 1}$ is the adjusted concept score. $f_p(\mathbf{D})$ is a pooling on the raw detection score matrix $\mathbf{D}$: each row corresponds to a shot and each column corresponds to a concept.

- Our goal is to generate video representations that tends to be similar to the underlying concept representation in terms of the distributional and logical consistency.

Concept Adjustment Model: Distributional Consistency

- Our general implementation:

$$g(\mathbf{v}; \alpha, \beta) = \alpha\beta\|\mathbf{v}\|_1 + (1-\alpha)\sum_{l=1}^{q}\beta\sqrt{p_l}\|\mathbf{v}^{(l)}\|_2,$$

  - When $\alpha = 1$ → concepts are independent.
  - When $\alpha = 0$ → groups of concepts frequently co-occur, e.g. sky/cloud, beach/ocean/waterfront, and table/chair. Multimodal concepts baby/baby_crying.
  - When $\alpha \in (0, 1)$ → only few concepts in a co-occurring group are nonzero [Simon et al. 2013].

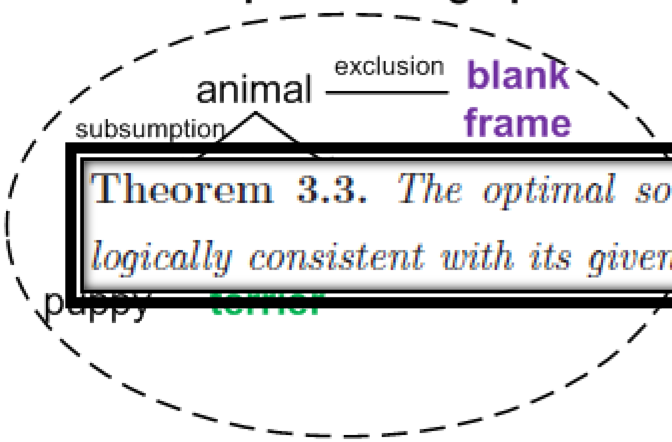
The choice of the model parameters depends on the underlying distribution of the semantic concepts in the dataset.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse group lasso. Journal of Computational and Graphical Statistics, 22(2):231–245,2013.

Concept Adjustment Model: Logical Consistency

Definition 3.1. A HEX graph $G = (N, E_h, E_e)$ is a graph consisting of a set of nodes $N = \{n_1, \cdots, n_m\}$, directed edges $E_h \subseteq N \times N$ and undirected edges $E_e \subseteq N \times N$ such that the subgraph $G_h = (N, E_h)$ is a directed acyclic graph and the subgraph $G_e = (N, E_e)$ has no self-loop. [Deng et al, 2014]

Concept relation graph



subsumption

$$v_{\text{dog}} \leq v_{\text{animal}}$$

Theorem 3.3. *The optimal solutions of Eq. (3.1) (before or after normalization) is logically consistent with its given HEX graph.*

$\leq \mathbf{c}$

exclusion

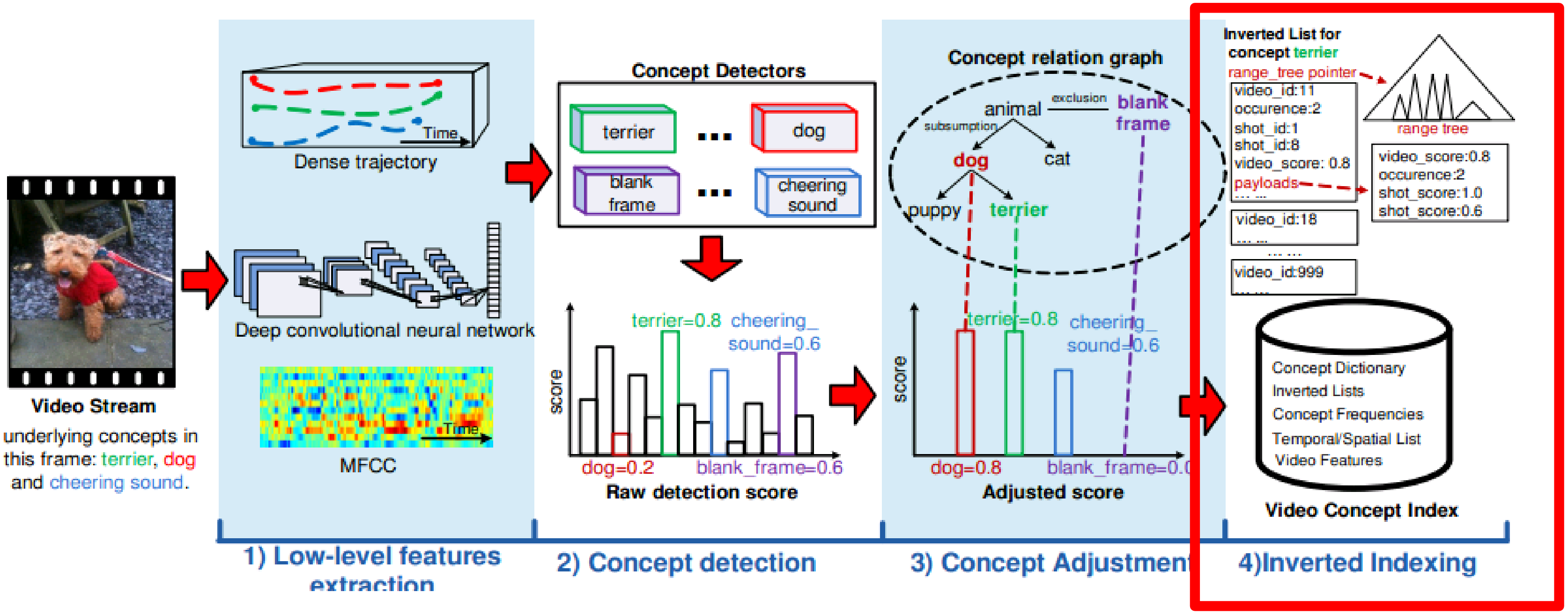
$$v_{\text{animal}} + v_{\text{blank\_frame}} \leq 1$$

$$v_{\text{animal}}, v_{\text{blank\_frame}} \in \{0, 1\}$$

Integer programming
solved by mix-integer toolbox or by constraint relaxation.

Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *ECCV, 2014.*

Indexing Semantic Features



- Finally, the adjusted concept representation is indexed by an inverted index. The index structure needs to be modified to account for:
  - Indexing real-valued concepts
  - Indexing the shot-level scores
  - Supporting Boolean logical and temporal operators.

Detailed methods are in Chapter 3

Experiments on MED

- Dataset: MED13Test and MED14Test (around 25,000 videos). Each set contains 20 events.
- Official evaluation metric: Mean Average Precision (MAP)
- Supplementary metrics:
  - Mean Reciprocal Rank = (1/rank of the first relevant sample)[Voorhees, 1999]
  - Precision@20
  - MAP@20
- Configurations:
  - NIST's HEX graph is used for IACC;
  - We build the HEX graphs for the rest of the semantic concept features.
  - Raw prediction scores of the 3000+ concepts trained in Chapter 7.

E.M. Voorhees. Proceedings of the 8th Text Retrieval Conference. TREC-8 Question Answering Track Report. 1999

Experiments on MED

Comparison of the raw and the adjusted representation

baseline

Method	Index	Evaluation Metric			
		P@20	MRR	MAP@20	MAP
MED13 Raw	385M	0.312	0.728	0.230	0.176
MED13 Adjusted	11.6M	0.325	0.689	0.247	0.172
MED14 Raw	357M	0.233	0.610	0.155	0.185
MED14 Adjusted	12M	0.219	0.540	0.144	0.171

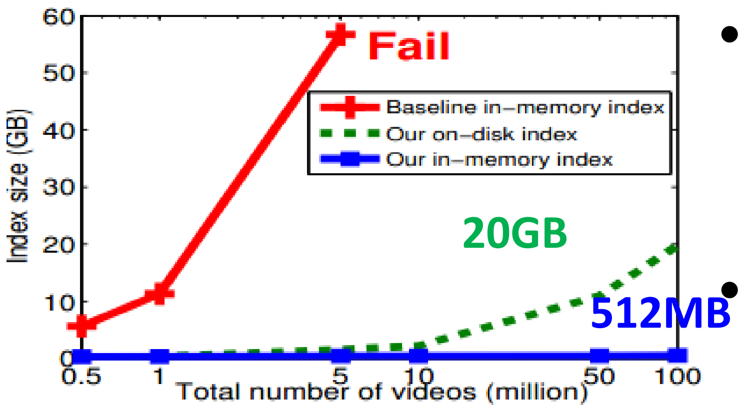
33x smaller index size

comparable performances

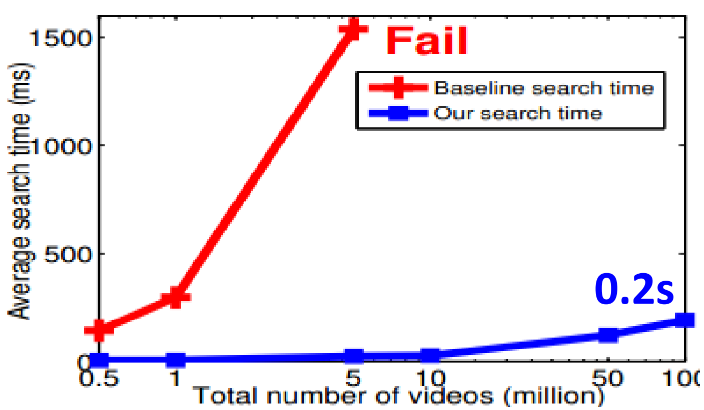
The accuracy of the proposed method is comparable to that of the baseline method.

Experiments on 100M Videos

The scalability and efficiency test on 100 million videos.



(a) Index (in GB)



(b) Search Time (in ms)

- Baseline method (raw score representation) fails when the data reaches 5 million videos.
- Our method can scale to 100M videos.
  - take 0.2s on a single core (on-line search time);
  - create an on-disk inverted index of 20G;
  - Use 512MB memory.

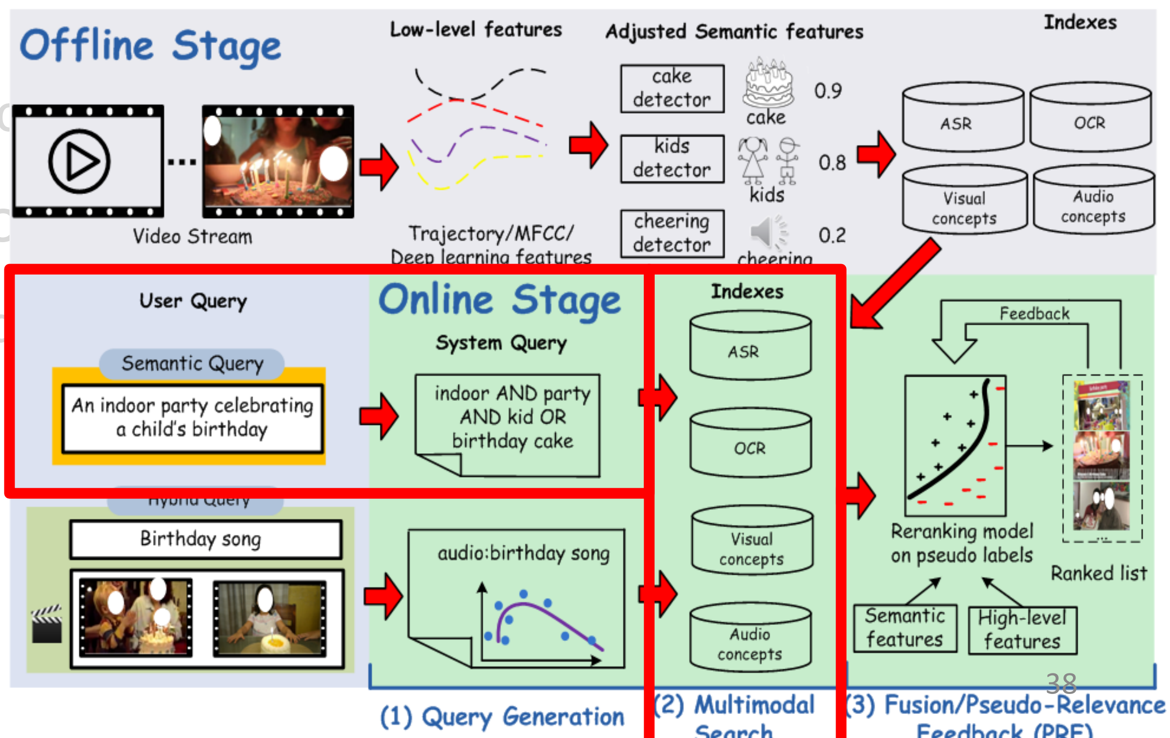
The proposed method is scalable and efficient.

Summary of Indexing Semantic Features

- We proposed a scalable semantic concept indexing methods that extends the current scale of video search by a few orders of magnitude while maintaining state-of-the-art retrieval performance.
- The key is a novel step called concept adjustment that can represent a video by a few salient and consistent concepts which can be efficiently indexed by a modified inverted index.

Outline

- Introduction
- Approaches:
  - Indexing Semantic Features [95%]
  - **Semantic Search [95%]** **: the search process for semantic queries.**
  - Video Reranking [95%]
  - Building Semantic Conc
- Conclusions and Propo
  - Proposed Work: Hybrid
    [10%]



Semantic Search:
Semantic Query Generation

- **(1) Semantic query generation**: how to map out-of-vocabulary query words to the concepts in our vocabulary?

user query

Making a sandwich

generated query

food, bread, cheese, kitchen, cooking, room, lunch, dinner;

  – The key is to measure the similarity between a query word and a concept in the vocabulary:
    - **Exact word matching**
    - **WordNet Similarity**: structural depths in WordNet taxonomy.
    - **Wikipedia Point-wise Mutual Information (PMI):** calculate the mutual information of two words in Wikipedia.
    - **Word embedding mapping**: word distance in a learned embedding space in Wikipedia by word2vec.

Semantic Query Generation

- We empirically study the following methods.

MAP comparison on MED13Test and MED14Test datasets

Individual methods are comparable.

Mapping Method	MAP 13Test	MAP 14Test	Time (s)
Exact Word Matching	9.66	7.22	0.10
WordNet	7.86	6.68	1.22
PMI	9.84	6.95	22.20
Word Embedding	8.79	6.21	0.48
Mapping Fusion	10.22	9.38	-

Fusion improves the mapping results

Semantic Search: Multimodal Search

- **(2) Retrieval methods**: what retrieval model to use for which modality?
  - Existing work [Dalton et al. 2013, Younessian et al 2012, Wu et al 2014] did not fully investigate the retrieval model's impact on multi-modalities.
  - We studied classical **four** retrieval models over **three** modalities: ASR, OCR, and semantic concepts
    - Vector Space Model (VSM): tf and tf-idf representations.
    - BM25
    - Language Model-JM Smoothing (LM-JM)
    - Language Model-Dirichlet Smoothing (LM-DL)
  - We found retrieval models have substantial impacts to the search result.
    - For ASR, **LM-JM** works the best. More than 1.5x better than the second best model.
    - For semantic concepts and OCR, **BM25** seems to be a robust and accurate retrieval model.

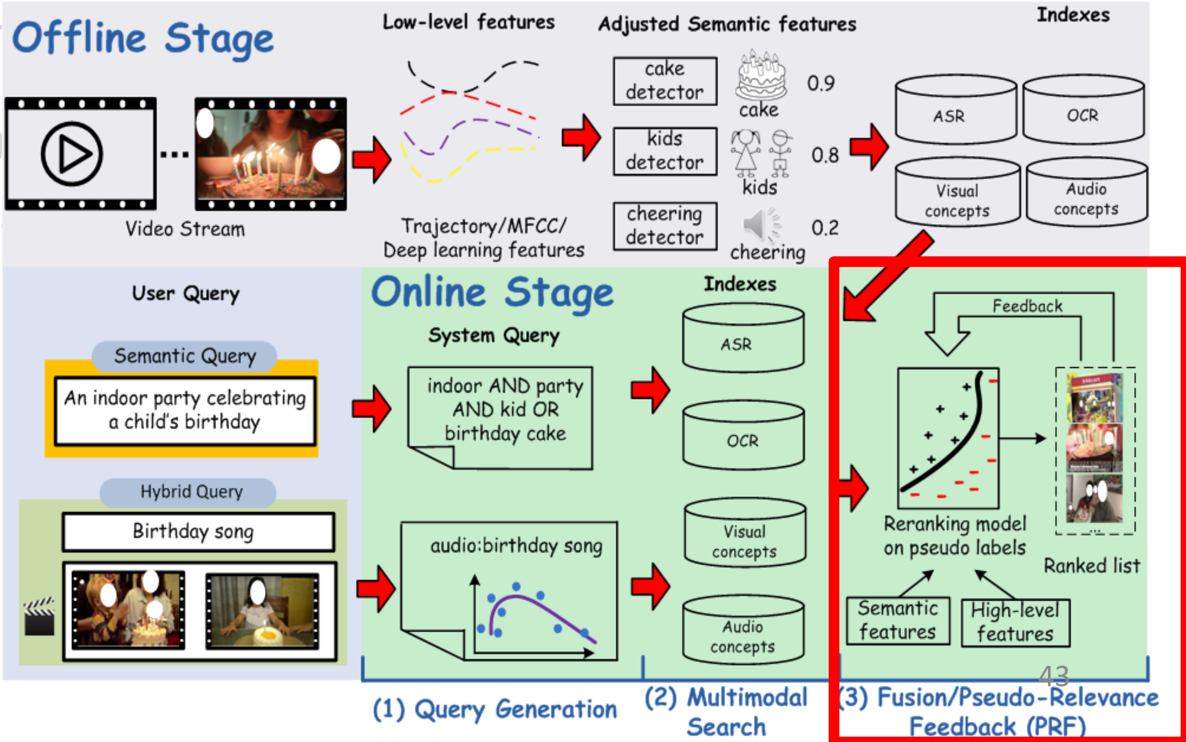
Ehsan Younessian, Teruko Mitamura, and Alexander Hauptmann. Multimodal knowledge-based analysis in multimedia event detection. In ICMR, 2012.
Jeffrey Dalton, James Allan, and Pranav Mirajkar. Zero-shot video retrieval using content and concepts. In CIKM, 2013.
Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In CVPR, 2014.

Summary of Semantic Search

- We empirically studied the semantic query generation and retrieval methods. We found that:
  - The fusion of mapping methods perform better than any individual methods.
  - Language Model-JM Smoothing works the best for ASR and BM25 works reasonably well for other types of features.

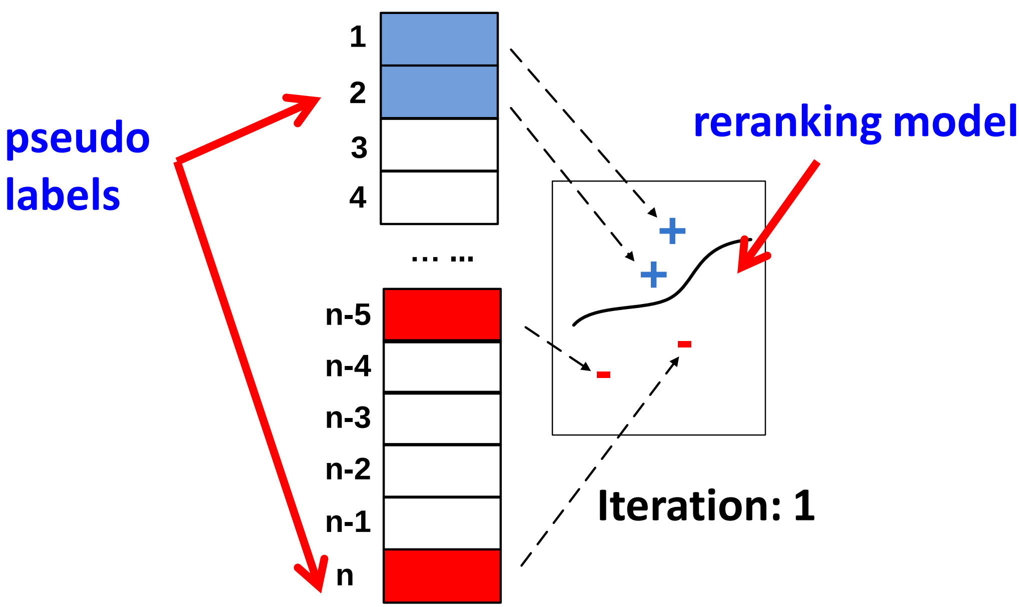
Outline

- Introduction
- Approaches:
  - Indexing Semantic Features [95%]
  - Semantic Search [95%]
  - **Video Reranking [95%]**
  - Building Semantic Cor
- Conclusions and Prop
  - Proposed Work: Hybr [10%]



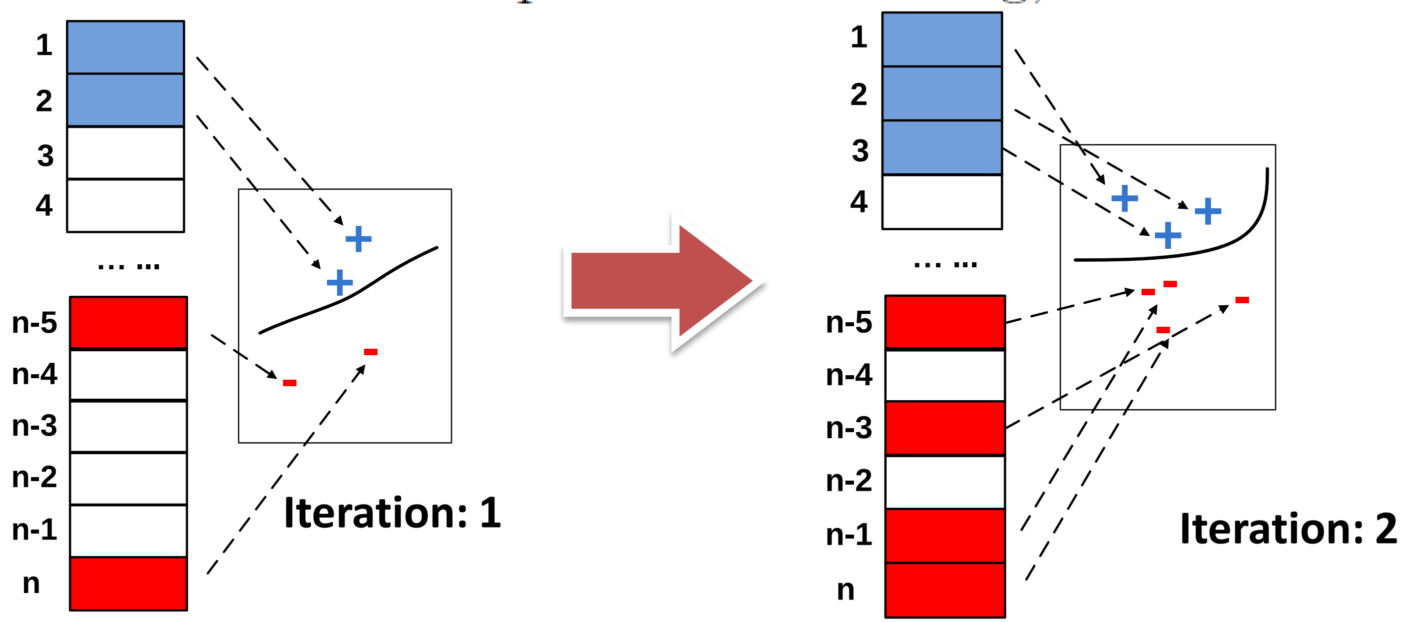
Generic Reranking Algorithm

```
t = 0; //Iteration zero
Choose the initial pseudo labels and weights;
while t ≤ max iteration do
    Train a reranking model on the fixed labels and weights;
    Update the pseudo labels and weights;
    if t is small then add more pseudo positives;
end while
return The list of samples after reranking;
```

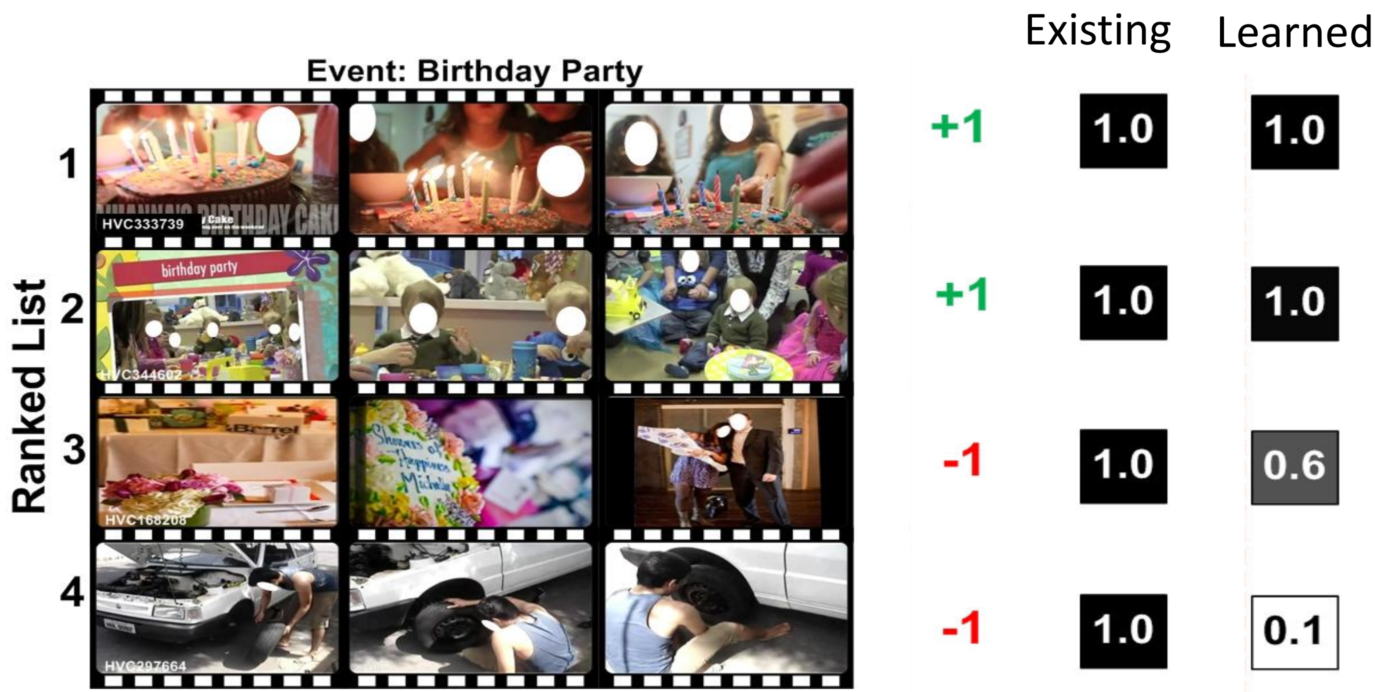


Generic Reranking Algorithm

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



Intuition



- 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**.

Self-paced Learning

- Self-paced learning (Kumar et al 2010) is a learning paradigm that is inspired by the learning process of humans and animals.
- The samples are not learned randomly but organized in **a meaningful order** which illustrates from easy to gradually more complex ones.

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, pages 1189–1197, 2010.

Self-paced Learning

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

Ranked list of iteration 1

Ranked List

1	+
2	+
3	
4	
5	
6	
7	
8	

… …

Age

Ranked list of iteration n

Ranked List

1	+
2	+
3	+
4	+
5	+
6	+
7	+
8	

… …

Self-paced Reranking (SPaR)

- The propose model:

$$\min_{\Theta_1,\dots,\Theta_m,\mathbf{y},\mathbf{v}} \mathbb{E}(\Theta_1,\dots,\Theta_m,\mathbf{v},\mathbf{y};C,k)$$

$$= \min_{\mathbf{y},\mathbf{v},\Theta_1,\dots,\Theta_m,} C\sum_{i=1}^{n} v_i \boxed{\sum_{j=1}^{m} \ell_{ij} + \sum_{j=1}^{m} \frac{1}{2}\|\mathbf{w}_j\|_2^2} + \boxed{mf(\mathbf{v};k)}$$

$$\text{s.t. } \forall i, \forall j, y_i(\mathbf{w}_j^T \phi(\mathbf{x}_{ij}) + b_j) \geq 1 - \ell_{ij}, \ell_{ij} \geq 0$$
$$\mathbf{y} \in \{-1,+1\}^n,$$
$$\mathbf{v} \in [0,1]^n,$$

$\Theta_1, \dots, \Theta_m$ Reranking models for each modality.

$\mathbf{y} \in \{-1,1\}^n$ The pseudo label.

$\mathbf{v} \in [0,1]^n$ The weight for each sample.

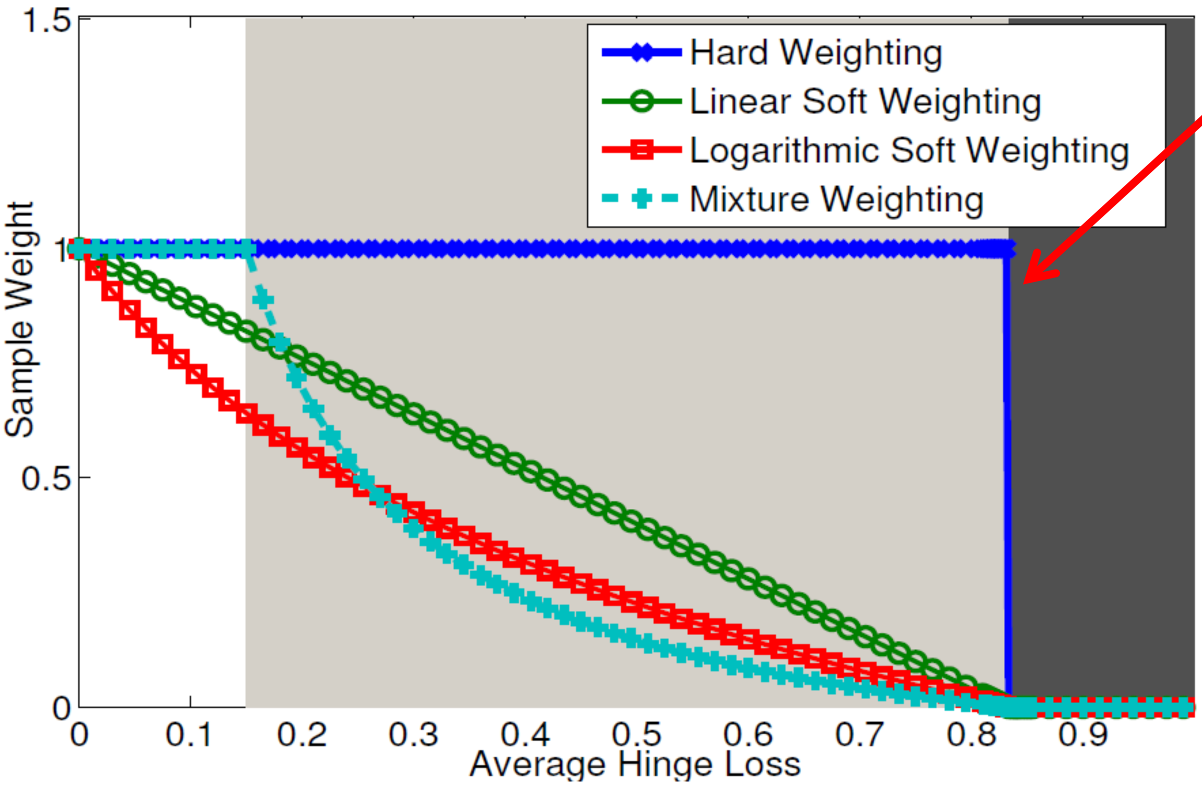
Hinge loss function

Function determines the weighting scheme

The self-paced is implemented by a regularizer.

The loss in the reranking model is discounted by a weight.

Proposed Weighting Schemes



Existing

Binary weighting [Kumar et al 2010]

$$f(\mathbf{v};k) = -\frac{1}{k}\|\mathbf{v}\|_1 = -\frac{1}{k}\sum_{i=1}^{n} v_i.$$

Proposed

linear weighting

$$f(\mathbf{v};k) = \frac{1}{k}\left(\frac{1}{2}\|\mathbf{v}\|_2^2 - \sum_{i=1}^{n} v_i\right).$$

Logarithmic weighting

$$f(\mathbf{v};k) = \sum_{i=1}^{n}\left(\zeta v_i - \frac{\zeta^{v_i}}{\log \zeta}\right),$$

Mixture weighting

$$f(\mathbf{v};k,k') = -\zeta \sum_{i=1}^{n} \log(v_i + \zeta k),$$

Reranking in Optimization and Conventional Perspective

$t = 0$; //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 $[v_1 y_1, \cdots, v_n y_n]^T$;

SPaR solution
Optimization perspective

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

Reranking solution
Conventional perspective

- Optimization perspective → theoretical justifications
- Conventional perspective → practical lessons

Q: Does the process converge? If so, to where?
A: For the proposed weighting, yes, to the local optimum.

Theorem 6.2. *The algorithm in Fig. 6.2 converges to a stationary solution for any fixed C and k.*

See the proof in Appendix D

Experiments on MED13Test

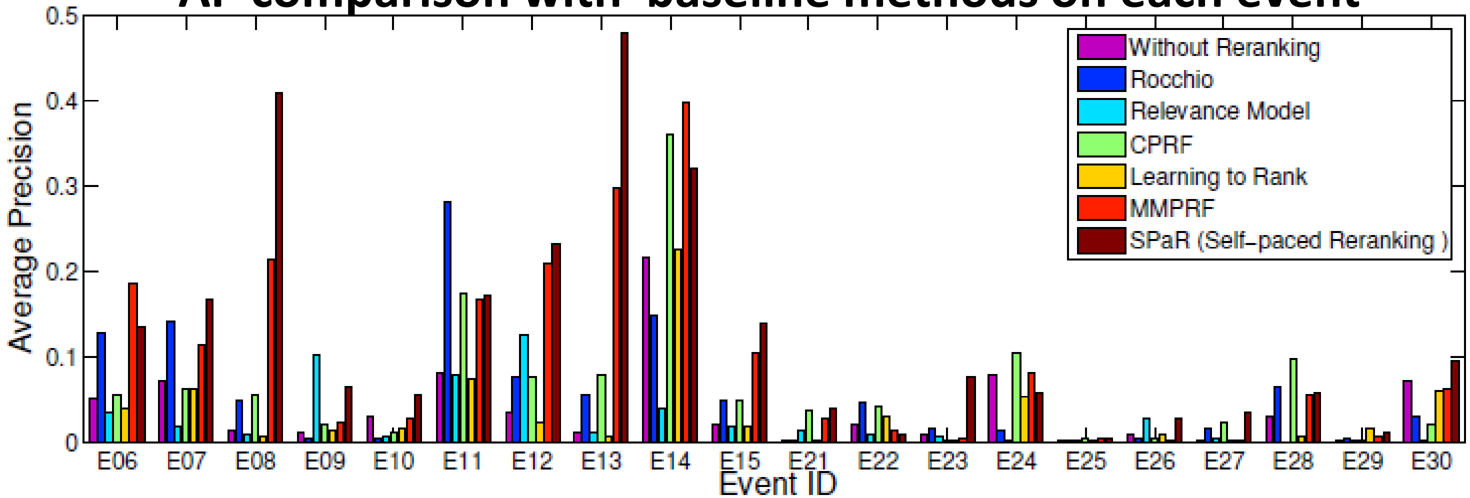
MAP (x100) comparison with baseline methods

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	**12.9**	$\mathbf{15.3 \pm 2.6}$

proposed method

Mixture weighting is used.

AP comparison with baseline methods on each event



Significant improvement!

Outperforms baseline methods on 15/20 events.

Experiments on Web Query

- Web image (353 queries over 71,478 images)
- Densely sampled SIFT are extracted.
- Parameters are tuned on a validation set.
- Mixture self-paced function is used.

MAP and MAP@100 comparison with baseline methods

Method	MAP	MAP@100
Without Reranking [17]	0.569	0.431
CPRF [38]	0.658	-
Random Walk [10]	0.616	-
Bayesian Reranking [33, 32]	0.658	0.529
Preference Learning Model [32]	-	0.534
BVLS [26]	0.670	-
Query-Relative(visual) [17]	0.649	-
Supervised Reranking [39]	0.665	-
SPaR	**0.672**	**0.557**

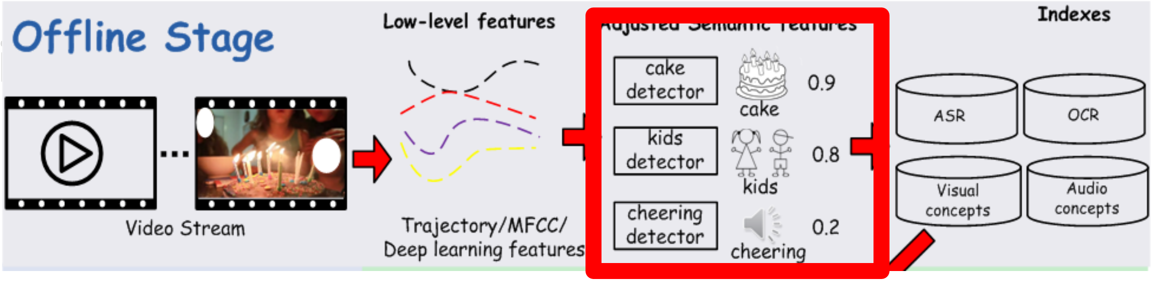
SPaR also works for image reranking (single modality)

Discussions on Video Reranking

- We proposed SPaR, a novel and general framework for multimodal reranking.
- It has theoretical justification, e.g. convergence properties.
- We found two scenarios where SPaR may fail:
  - Initial top-ranked samples are completely off-topic (bad starting values).
  - Features used in reranking are not discriminative to the queries.

Outline

- Introduction
- Approaches:
  - Indexing Semantic Features [95%]
  - Semantic Search [95%]
  - Video Reranking [95%]
  - **Building Semantic Concepts [80%]**
- Conclusions and Proposed Work
  - Proposed Work: Hybri
    [10%]



Introduction: Building Semantic Concepts

- Training concept detectors need lots of labeled training data. Annotated video data are hard to collect.
- Our solution is to train detectors from weakly labeled video data (metadata) downloaded from the Internet.
  - Pros: no manual annotations
  - Cons: weakly labeled data are very noisy
- We are interested in approaching this problem in a more principled and theoretically sound way.
  - Derive a theory from paradigms of curriculum learning and self-paced learning.
  - Use proposed theory to train concept detectors on noisy data.

Curriculum Learning and Self-paced Learning

Learning philosophy[Bengio et al. 2009, Kumar et al. 2010]:

- Learning is an iterative process.
- Samples should be organized in a meaningful order (**called curriculum**).
- Model complexity increases in each iteration.

“bus” to learn earlier

Age →

“bus” to learn later

*The above of real examples in the TRECVID SIN dataset (http://trecvid.nist.gov/).

Curriculum Learning and Self-paced Learning

- **Curriculum Learning (CL)**: assign learning priorities to training samples, according to prior knowledge or heuristics about specific problems [Bengio et al. 2009].
  - parsing from shorter sentences to longer sentence [Spitkovsky et al. 2009].
- **Self-paced Learning (SPL):** the curriculum is determined by the learned models. Solving a joint optimization problem of the learning objective with the latent curriculum [Kumar, Packer, and Koller 2010].
  - Broadly used in many learning problems such as tracking[Supancicet al. 2013], domain adaptation [Tang et al. 2012], segmentation [Kumar et al. 2011], etc.

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, pages 1189–1197, 2010.
Kevin Tang, Vignesh Ramanathan, Li Fei-Fei, and Daphne Koller. Shifting weights: Adapting object detectors from image to video. In NIPS, 2012
M. Kumar, H. Turki, D. Preston, and D. Koller. Learning specific-class segmentation from diverse data. In ICCV, 2011.
J. Supanˇciˇc III and D. Ramanan. Self-paced learning for long-term tracking. In CVPR, 2013.
V. I. Spitkovsky, H. Alshawi, and D. Jurafsky. Baby steps: How less is more in unsupervised dependency parsing. In NIPS, 2009.

Curriculum Learning versus Self-paced Learning

Curriculum Learning (CL)

- Pros
  - Flexible to incorporate prior knowledge/heuristics.
- Cons
  - Curriculum is determined beforehand which may not be consistent with dynamically learned models.

Self-paced Learning (SPL)

- Pros
  - Learn consistent models.
  - Concise optimization problem.
- Cons
  - Cannot use prior knowledge.
  - Random starting values (can significantly affect performance).

Unified in a single framework:
Self-paced Curriculum Learning

Self-paced Learning

- Formulated as an optimization problem (based on SPL).

Learner

$$\arg \min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \sum_{i=1}^{n} v_i \boxed{L(y_i, g(\mathbf{x}_i, \mathbf{w}))} - \lambda \sum_{i=1}^{n} v_i$$

$$\left.\begin{aligned} \mathbf{w} &\Rightarrow \text{parameters in the off-the-shell model} \\ L(y_i, g(\mathbf{x}_i, w)) &\Rightarrow \text{loss for the } ith \text{ sample} \end{aligned}\right\}$$

Off-the-shelf model (SVM, neural networks etc.)

$$\mathbf{v} = [v_1, \ldots, v_n] \Rightarrow \text{latent weight vector for all samples}$$

- While fixing w, the solution is:

$$v_i^* = \begin{cases} 1, & L(y_i, g(\mathbf{x}_i, \mathbf{w})) < \lambda, \\ 0, & \text{otherwise.} \end{cases} \quad \lambda \Rightarrow \text{model age}$$

Self-paced Curriculum Learning

- Proposed learning objectives:

Learning schemes

$$\arg \min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \sum_{i=1}^{n} v_i L(y_i, g(\mathbf{x}_i, \mathbf{w})) + \boxed{f(\mathbf{v}, \lambda)}$$

$$\text{subject to } \mathbf{v} \in \Psi$$

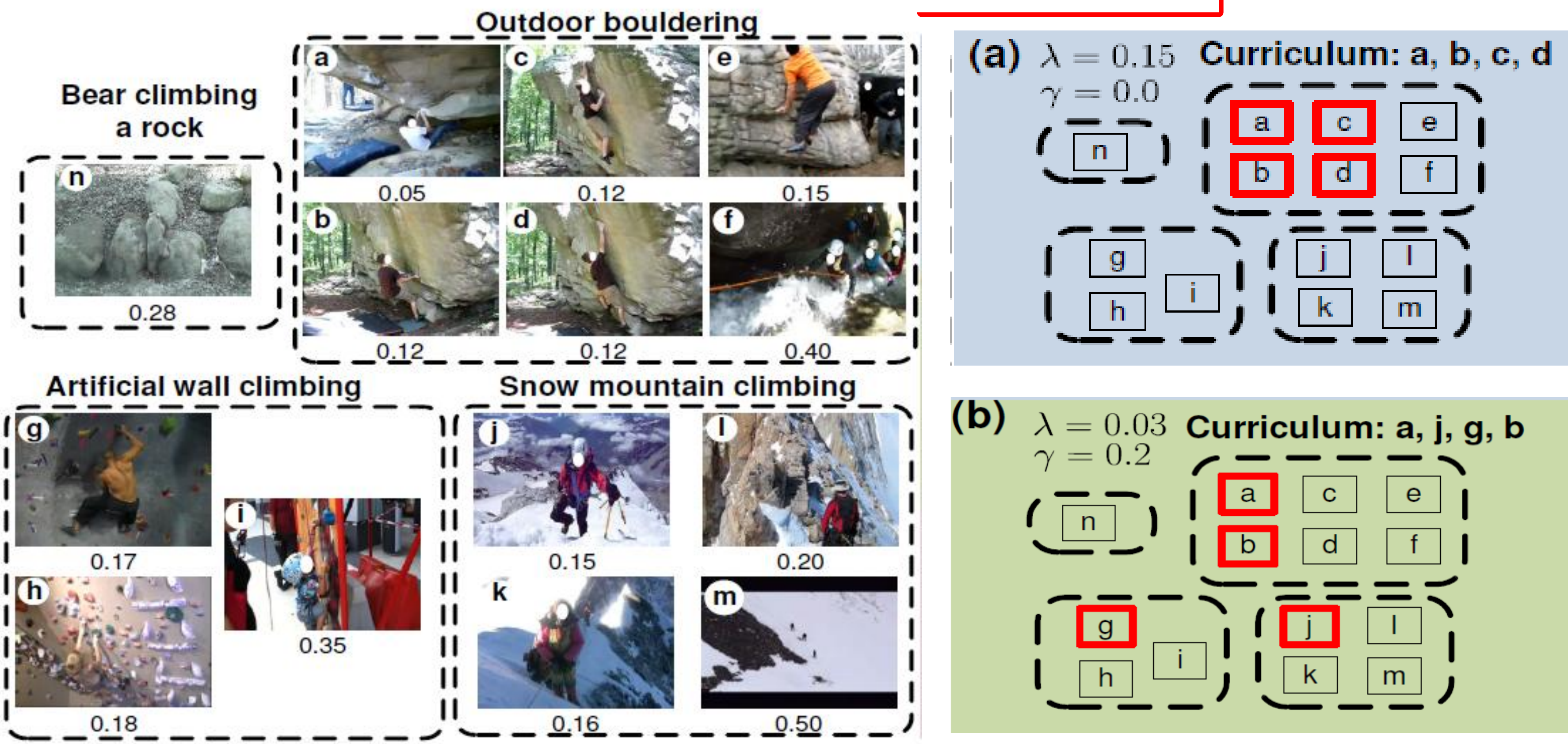
$f(\mathbf{v}, \lambda) \Rightarrow$ regularizer determines the learning scheme

Generalize a single learning scheme to multiple learning schemes.
For different problems, we can use different learning schemes.

“Rock Climbing”
Learning Easy and Diverse Samples

$$f(\mathbf{v}, \lambda) = -\lambda \sum_{i=1}^{n} v_i$$

Favor diverse examples



Self-paced Curriculum Learning

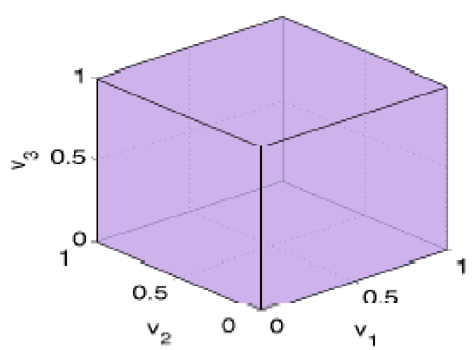
- Proposed learning objectives:

$$\arg\min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \sum_{i=1}^{n} v_i L(y_i, g(\mathbf{x}_i, \mathbf{w})) + f(\mathbf{v}, \lambda)$$

$$\text{subject to } \mathbf{v} \in \Psi$$

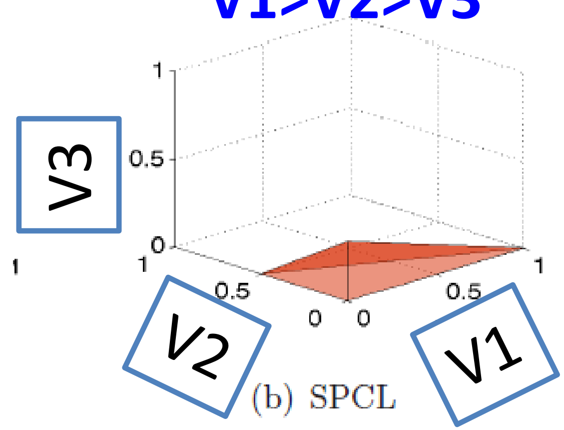
Prior knowledge in curriculum learning

- The shape of the feasible region weakly implies a prior learning sequence of samples.



(a) SPL

V1>V2>V3



(b) SPCL

Self-paced Curriculum Learning

- Proposed learning objectives:

Learner

Learning schemes

$$\arg \min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \sum_{i=1}^{n} v_i L(y_i, g(\mathbf{x}_i, \mathbf{w})) + f(\mathbf{v}, \lambda)$$

$$\text{subject to } \mathbf{v} \in \Psi$$

Prior knowledge in curriculum learning

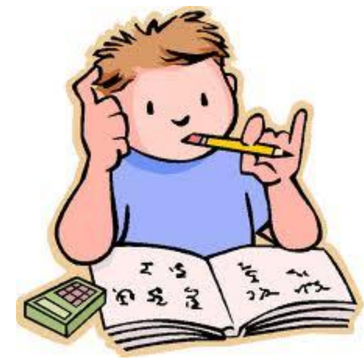
- A new learning theory:
  - Flexible learning schemes to fit various problems;
  - Easy to incorporate prior knowledge;
  - Support any loss function.

Self-paced Curriculum Learning

Curriculum Learning (CL)

instructor-driven

Self-paced Learning (SPL)



student-driven

Self-paced Curriculum Learning (SPCL)

instructor-student-collaborative

Unified in a single framework: SPCL

Preliminary Experiments

Comparison of SPL and SPCL with diversity learning scheme on MED

Run Name	RandomForest	AdaBoost	BatchTrain	SPL	SPLD
Best Run	3.0	2.8	8.3	9.6	**12.1**
10 Runs Average	3.0	2.8	8.3	8.6 ± 0.42	$\mathbf{9.8 \pm 0.45}$

Proposed method

Comparison of SPL and SPCL with diversity learning scheme on Hollywood2 and Olympic Sports

Run Name	RandomForest	AdaBoost	BatchTrain	SPL	SPLD
Hollywood2	28.20	41.14	58.16	63.72	**66.65**
Olympic Sports	63.32	69.25	90.61	90.83	**93.11**

Proposed method

See more experiments in Section 7.4

Preliminary Experiments

- Using the proposed theory, we build detectors using the YFCC videos (videos sampled from Flickr) with no labels.
- We derive the curriculum from metadata (using language models) and train SPCL with diversity learning scheme.
- Train 609 detectors over 400K weakly labeled videos.
- We manually evaluate their P@10 on a third dataset (MED).

YFCC 609	ImageNet 1000	UCF 101
0.37608	0.2063	<0.1

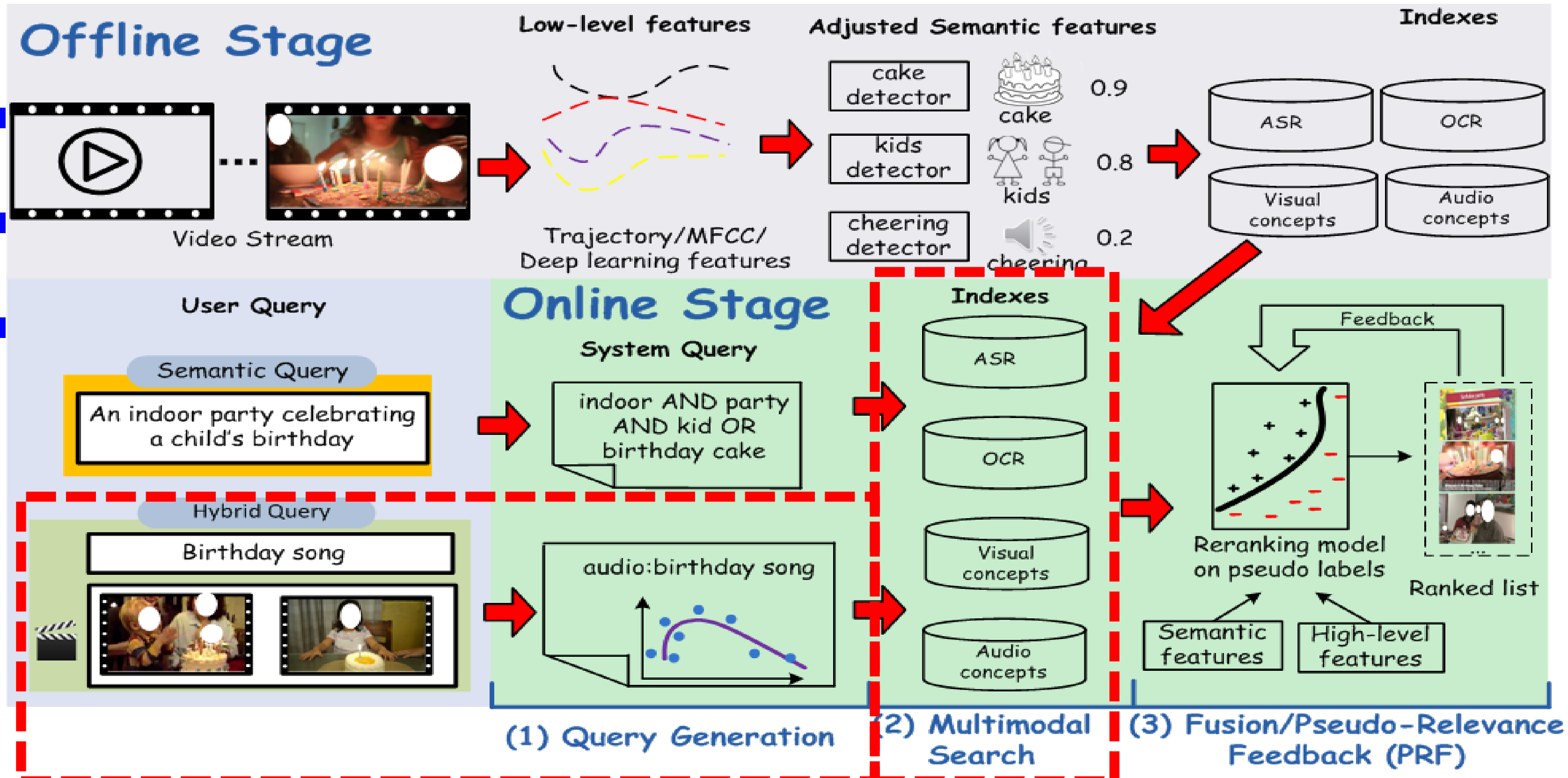
10k labeled video segments

Weakly labels on 400K videos

1M labeled still images

Detectors built on a large weakly labeled data set are more accurate than those built on a small labeled dataset.

Outline



- **Conclusions**
  - **Proposed Work: hybrid search**

Proposed Work

- Processing hybrid queries:
  - Preliminary studies showed hybrid query with 10 examples can be done efficiently on compressed semantic features.
    - Shoou-I Yu, Lu Jiang, Zhongwen Xu, Yi Yang, Alexander Hauptmann. Content-Based Video Search over 1 Million Videos with 1 Core in 1 Second. In ACM International Conference on Multimedia Retrieval (ICMR), 2015.
  - The method, however, is not scalable as it needs preloading lots of data into the memory.
  - We plan to integrate semantic search methods into hybrid search
    - Use the compressed semantic features.
    - Apply concept adjustment.
    - Apply semantic search to filter out irrelevant samples.
  - We will test the proposed methods on MED and YFCC datasets.
- Training concept detectors on the whole YFCC dataset (about 0.8 million videos.)

Schedule

- October – Jan, 2015. Study the efficient search model for hybrid search.
- February – March, 2016. Test the model and finish the experiments.
- April – September, 2016. Thesis writing and defense.

Published papers on the thesis topic

[MM15] Lu Jiang, Shoou-I Yu, Deyu Meng, Yi Yang, Teruko Mitamura, Alexander Hauptmann. Fast and Accurate Content-based Semantic Search in 100M Internet Videos. In ACM Multimedia (MM), 2015.

[ICMR15] Lu Jiang, Shoou-I Yu, Deyu Meng, Teruko Mitamura, Alexander Hauptmann. Bridging the Ultimate Semantic Gap: A Semantic Search Engine for Internet Videos. In ACM International Conference on Multimedia Retrieval (ICMR), 2015. **[best paper candidate]**

[AAAI15] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, Alexander Hauptmann. Self-paced Curriculum Learning. In Conference on Artificial Intelligence (AAAI), 2015.

[NIPS14] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhen-Zhong Lan, Shiguang Shan, Alexander Hauptmann. Self-paced Learning with Diversity. In Neural Information Processing Systems (NIPS), 2014.

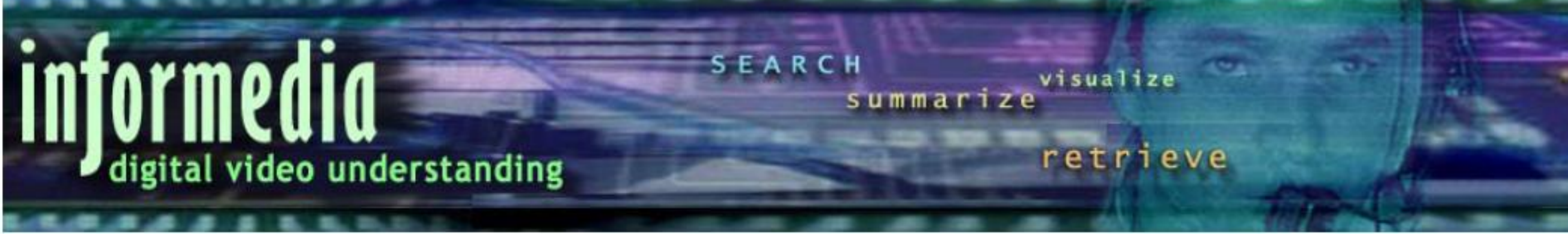
[MM14] Lu Jiang, Deyu Meng, Teruko Mitamura, Alexander Hauptmann. Easy Samples First: Selfpaced Reranking for Zero-Example Multimedia Search. In ACM Multimedia (MM), 2014.

[ICMR14] Lu Jiang, Teruko Mitamura, Shoou-I Yu, Alexander Hauptmann. Zero-Example Event Search using MultiModal Pseudo Relevance Feedback. In ACM International Conference on Multimedia Retrieval (ICMR), 2014.

[ICMR14] Lu Jiang, Wei Tong, Deyu Meng, Alexander Hauptmann. Towards Efficient Learning of Optimal Spatial Bag-of-Words Representations. In ACM International Conference on Multimedia Retrieval (ICMR). 2014. **[best paper candidate]**

[SLT14] Yajie Miao, Lu Jiang, Hao Zhang, Florian Metze. Improvements to Speaker Adaptive Training of Deep Neural Networks. In IEEE Spoken Language Technology (SLT), 2014. **[best poster]**

[MM12] Lu Jiang, Alexander Hauptmann, Guang Xiang. Leveraging High-level and Low-level Features for Multimedia Event Detection. In ACM Multimedia (MM), 2012.



Key Contributions:

- The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR'15]. The proposed framework supports text-to-video, video-to-video, and text&video-to-video search [MM'12].

- A novel theory about self-paced curriculums learning and its application on robust concept detector training [NIPS'14, AAAI'15].

- Novel reranking algorithms for improving performance [MM'14, ICMR'14].

- A concept adjustment method representing a video by a few salient and consistent concepts that can be efficiently indexed by the modified inverted index [MM'15]

THANK YOU.
QUESTIONS?

References

- Baptist Vandersmissen, Fr´ederic Godin, Abhineshwar Tomar, Wesley De Neve,and Rik Van de Walle. The rise of mobile and social short-form video: an indepth measurement study of vine. In ICMR Workshop on Social Multimedia and Storytelling, 2014.
- Masoud Mazloom, Xirong Li, and Cees GM Snoek. Few-example video event retrieval using tag propagation. In ICMR, 2014.
- Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In CVPR, 2014.
- Hyungtae Lee. Analyzing complex events and human actions in" in-the-wild" videos. In UMD Ph.D Theses and Dissertations, 2014.
- Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparsegroup lasso. Journal of Computational and Graphical Statistics, 22(2):231–245,2013.
- Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In ECCV, 2014.
- E.M. Voorhees. Proceedings of the 8th Text Retrieval Conference. TREC-8 Question Answering Track Report. 1999
- Ehsan Younessian, Teruko Mitamura, and Alexander Hauptmann. Multimodal knowledge-based analysis in multimedia event detection. In ICMR, 2012.
- Jeffrey Dalton, James Allan, and Pranav Mirajkar. Zero-shot video retrieval using content and concepts. In CIKM, 2013.
- Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In CVPR, 2014.
- R. Yan, A. G. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In CVIR, 2003.

References

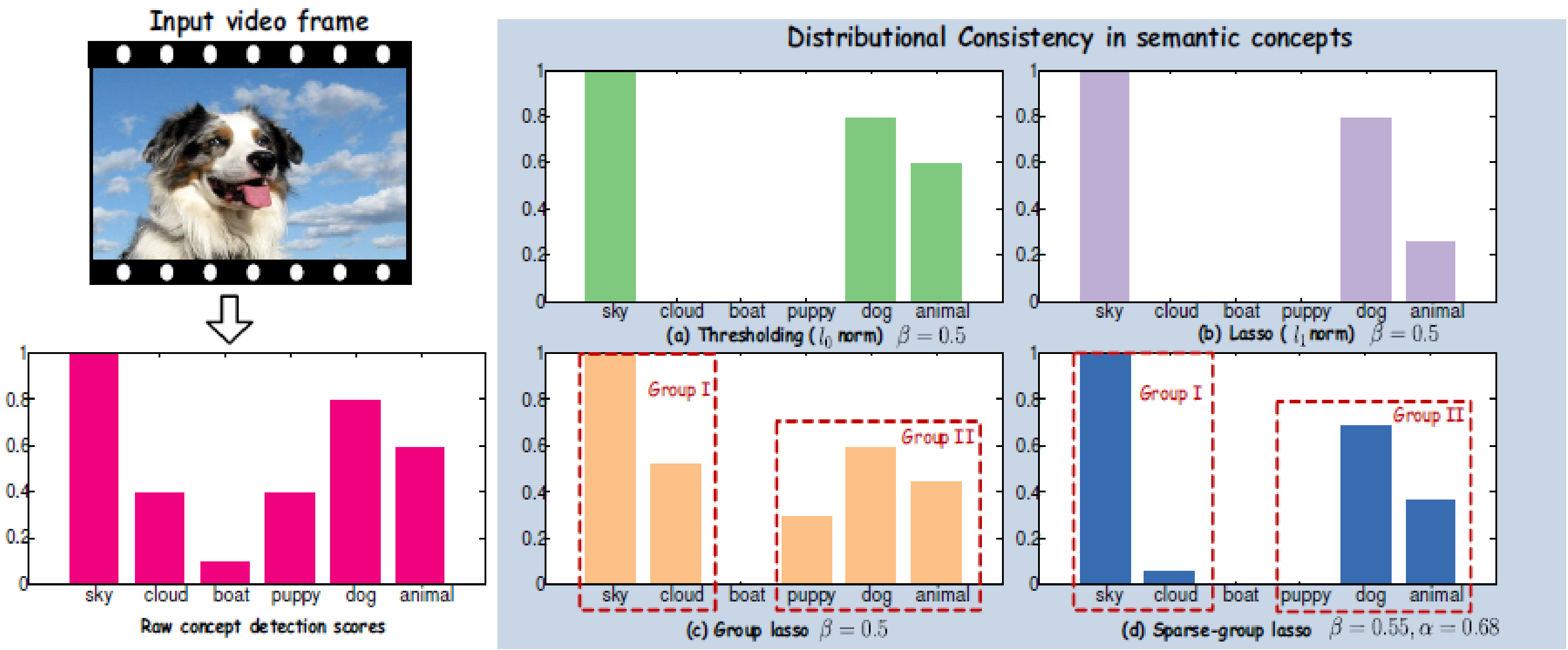
- A. G. Hauptmann, M. G. Christel, and R. Yan. Video retrieval based on semantic concepts. Proceedings of the IEEE, 96(4):602–622, 2008.
- L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In ICMR, 2014
- W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. In Multimedia, 2007.
- Y. Liu, T. Mei, X.-S. Hua, J. Tang, X. Wu, and S. Li. Learning to video search rerank via pseudo preference feedback. In ICME, 2008.
- X. Tian, Y. Lu, L. Yang, and Q. Tian. Learning to judge image search results. In Multimedia, 2011.
- X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In Multimedia, 2008.
- W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. In Multimedia, 2007.
- L. Nie, S. Yan, M. Wang, R. Hong, and T.-S. Chua. Harvesting visual concepts for image search with complex queries. In Multimedia, 2012.
- 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, pages 1189–1197, 2010.
- Kevin Tang, Vignesh Ramanathan, Li Fei-Fei, and Daphne Koller. Shifting weights: Adapting object detectors from image to video. In NIPS, 2012.
- M. Kumar, H. Turki, D. Preston, and D. Koller. Learning specific-class segmentation from diverse data. In ICCV, 2011.
- J. Supanˇciˇc III and D. Ramanan. Self-paced learning for long-term tracking. In CVPR, 2013.
- V. I. Spitkovsky, H. Alshawi, and D. Jurafsky. Baby steps: How less is more in unsupervised dependency parsing. In NIPS, 2009.

APPENDIX

Applications

- It can benefit a variety of related tasks such as video summarization [7], video recommendation, video hyperlinking [8], social media video stream analysis [9], in-video advertising [10], etc.

Distributional Consistency: A Toy Example



Experiments on MED

Comparison of the full adjustment model with its special case Top-*k* Thresholding on using IACC features.

Method	k	Evaluation Metric P@20	MRR	MAP@20	MAP
Our Model	50	**0.0392**	**0.137**	**0.0151**	**0.0225**
Top-k	50	0.0342	0.0986	0.0117	0.0218
Our Model	60	**0.0388**	**0.132**	**0.0158**	**0.0239**
Top-k	60	0.0310	0.103	0.0113	0.0220

Example Queries

- Using the query, our system should be able to
  - Find simple objects, actions, speech words.
  - Search complex activities.
  - Answer questions by/in videos.

Information need:
What did we talk about in the last year's forest camp?

Query (search videos in last year):
forest
AND (walking OR hiking)
OR tree
AND faces
AND asr:speech != empty

Concept Adjustment Model: Distributional Consistency

- A naive implementation → infeasible to solve.

$$g(\mathbf{v}; \alpha, \beta) = \frac{1}{2}\beta^2 \|v\|_0$$

- Our general implementation:

$$g(\mathbf{v}; \alpha, \beta) = \alpha\beta\|\mathbf{v}\|_1 + (1-\alpha)\sum_{l=1}^{q} \beta\sqrt{p_l}\|\mathbf{v}^{(l)}\|_2,$$

  - When $\alpha = 1$ → lasso (approximate l_0 norm).
  - When $\alpha = 0$ → group lasso (nonzero entries in a sparse set of groups)
  - When $\alpha \in (0, 1)$ → sparse group lasso (group-wise sparse solution, but only few coefficients in the group will be nonzero)

Experiments on YFCC

- We manually created queries for 30 products.
- Put commercials about the product to related video (in-video ads.)
- Evaluate the relevance of the top 20 returned results.

Average pe… nercials on YFCC

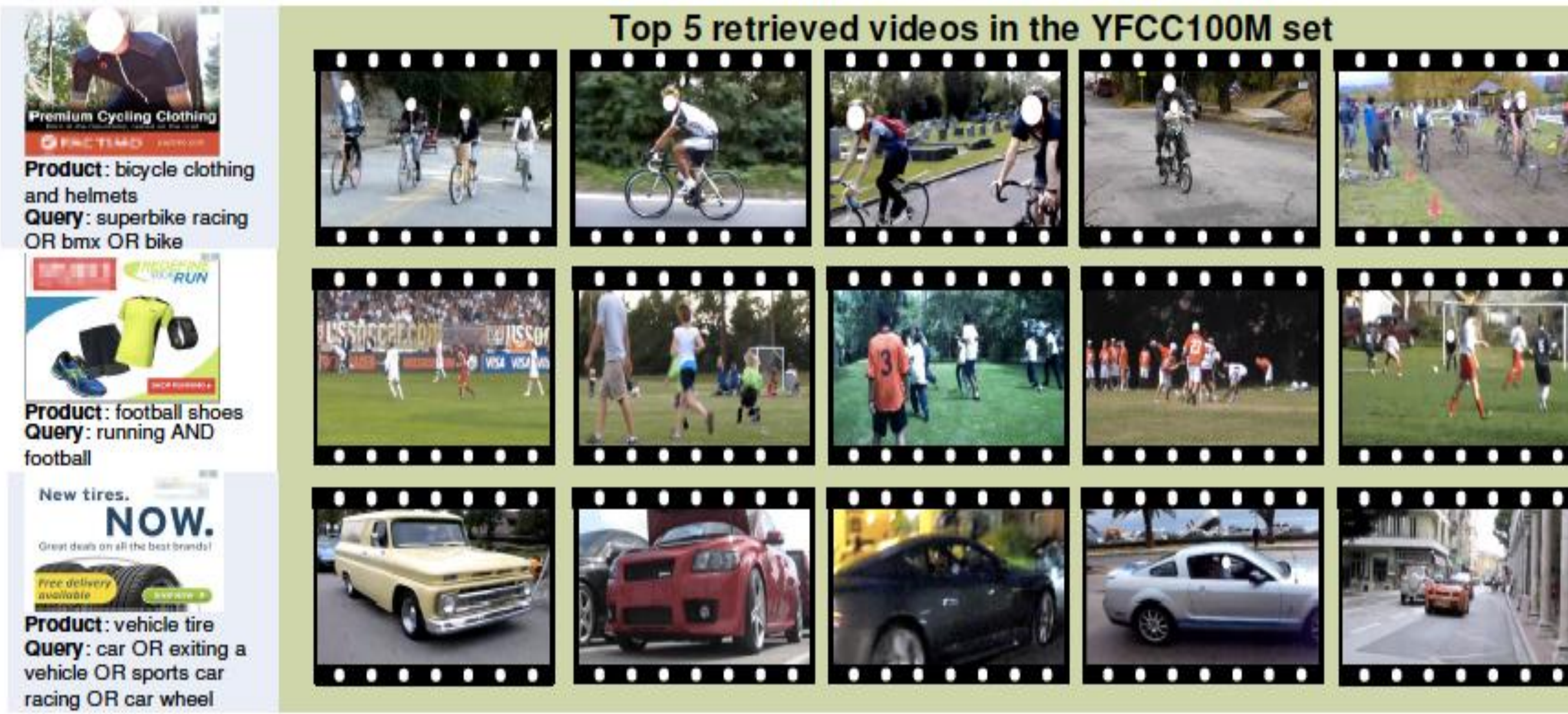
Category			ation Metric	
			IRR	MAP@20
Sports			1.00	0.94
Auto			1.00	0.95
Grocery			0.93	0.88
Traveling			1.00	0.96
Miscellan…			0.85	0.74
Average	30	0.81	0.93	0.86

Product: bicycle clothing and helmets
Query: superbike racing OR bmx OR bike

Queries and more results are available at:

https://sites.google.com/site/videosearch100m/

Experiments on YFCC



Self-paced Reranking (SPaR)

- The propose model:

$$\min_{\Theta_1,\dots,\Theta_m,\mathbf{y},\mathbf{v}} \mathbb{E}(\Theta_1, \dots, \Theta_m, \mathbf{v}, \mathbf{y}; C, k)$$

$$= \min_{\substack{\mathbf{y},\mathbf{v},\mathbf{w}_1,\dots,\mathbf{w}_m,\\ b_1,\dots,b_m,\{\ell_{ij}\}}} C\sum_{i=1}^{n} v_i \sum_{j=1}^{m} \ell_{ij} + \sum_{j=1}^{m} \frac{1}{2}\|\mathbf{w}_j\|_2^2 + \boxed{\text{regularizer}}$$

$$\text{s.t. } \forall i, \forall j, y_i(\mathbf{w}_j^T\phi(\mathbf{x}_{ij}) + b_j) \geq 1 - \ell_{ij}, \ell_{ij} \geq 0$$

$$\mathbf{y} \in \{-1, +1\}^n,$$

$$\mathbf{v} \in [0, 1]^n,$$

$\Theta_1, \dots, \Theta_m$ Reranking models for each modality.

$\mathbf{y} \in \{-1, 1\}^n$ The pseudo label.

$\mathbf{v} \in [0, 1]^n$ The weight for each sample.

For example the Loss in the SVM model.

$$\ell_{ij} = \max\{0, 1 - y_i \cdot (\mathbf{w}_j^T\phi(\mathbf{x}_{ij}) + b_j)\}$$

Self-paced Reranking (SPaR)

- The propose model:

$$\min_{\Theta_1,\dots,\Theta_m,\mathbf{y},\mathbf{v}} \mathbb{E}(\Theta_1,\dots,\Theta_m,\mathbf{v},\mathbf{y};C,k)$$

$$= \min_{\mathbf{y},\mathbf{v},\Theta_1,\dots,\Theta_m,} C\sum_{i=1}^{n} v_i \boxed{\text{loss-function}} + mf(\mathbf{v};k)$$

$$\text{s.t.} \boxed{\text{constraints}}$$

$$\mathbf{y} \in \{-1,+1\}^n,$$
$$\mathbf{v} \in [0,1]^n,$$

$\Theta_1,\dots,\Theta_m$ Reranking models for each modality.

$\mathbf{y} \in \{-1,1\}^n$ The pseudo label.

$\mathbf{v} \in [0,1]^n$ The weight for each sample.

The self-paced is implemented by a regularizer.
Physically corresponds to learning schemes that human use to learn different tasks.

m is the total number of modality.

f is the self-paced function in self-paced learning.

Reranking in Optimization and Conventional Perspective

```
t = 0; //Iteration zero
Choose starting values for y, v;
while t ≤ max iteration do
    Θ_1^(t+1), ..., Θ_m^(t+1) = arg max E_{y,v}(Θ_1^(t), ..., Θ_m^(t); C);
    y^(t+1), v^(t+1) = arg max E_Θ(y^(t), v^(t); k);
    if t is small then increase 1/k;
end while
return [v_1 y_1, ··· , v_n y_n]^T;
```

Optimization perspective

```
t = 0; //Iteration zero
Choose the initial pseudo labels and weights;
while t ≤ max iteration do
    Train a reranking model on the fixed labels and weights;
    Update the pseudo labels and weights;
    if t is small then add more pseudo positives;
end while
return The list of samples after reranking;
```

Conventional perspective

Q1: Why the reranking algorithm performs iteratively?

A: Self-paced learning mimicking human and animal learning process (from easy to complex examples).

Q2: Does the process converge? If so, to where?

A: Yes, to the local optimum. See the theorem in our paper.

Q3: Does the arbitrarily predefined weighting scheme converge?

A: No, but the weights by self-paced function guarantees the convergence.

Reranking in Optimization and Conventional Perspective

$t = 0$; //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 $[v_1 y_1, \cdots, v_n y_n]^T$;

Optimization perspective

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

Conventional perspective

Q1: Why the reranking algorithm performs iteratively?

A: Self-paced learning mimicking human and animal learning process (from easy to complex examples).

Q2: Does the process converge? If so, to where?

A: Yes, to the local optimum.

Q3: Does the arbitrarily predefined weighting scheme converge?

A: No, but the weights by self-paced function guarantees the convergence.

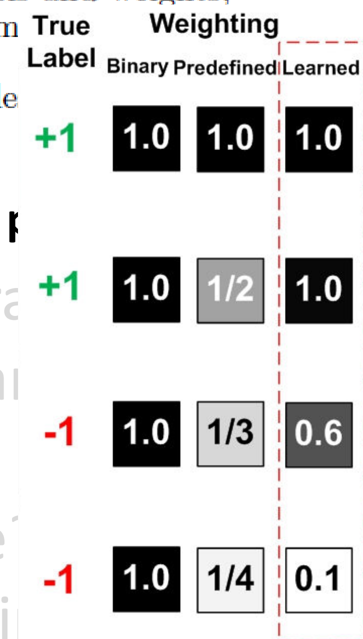
Reranking in Optimization and Conventional Perspective

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

Optimization perspective

$t = 0$; //Iteration zero
Choose the initial pseudo labels and weights;
while $t \leq$ max iteration **do**
$\quad$ Train a reranking model on the fixed labels and weights;
$\quad$ Update the pseudo labels and weights;
$\quad$ **if** t is small **then** add m
end while
return The list of sample

Conventional p



Q1: Why the reranking algorithm performs itera

A: Self-paced learning mimicking human and a g
process (from easy to complex examples).

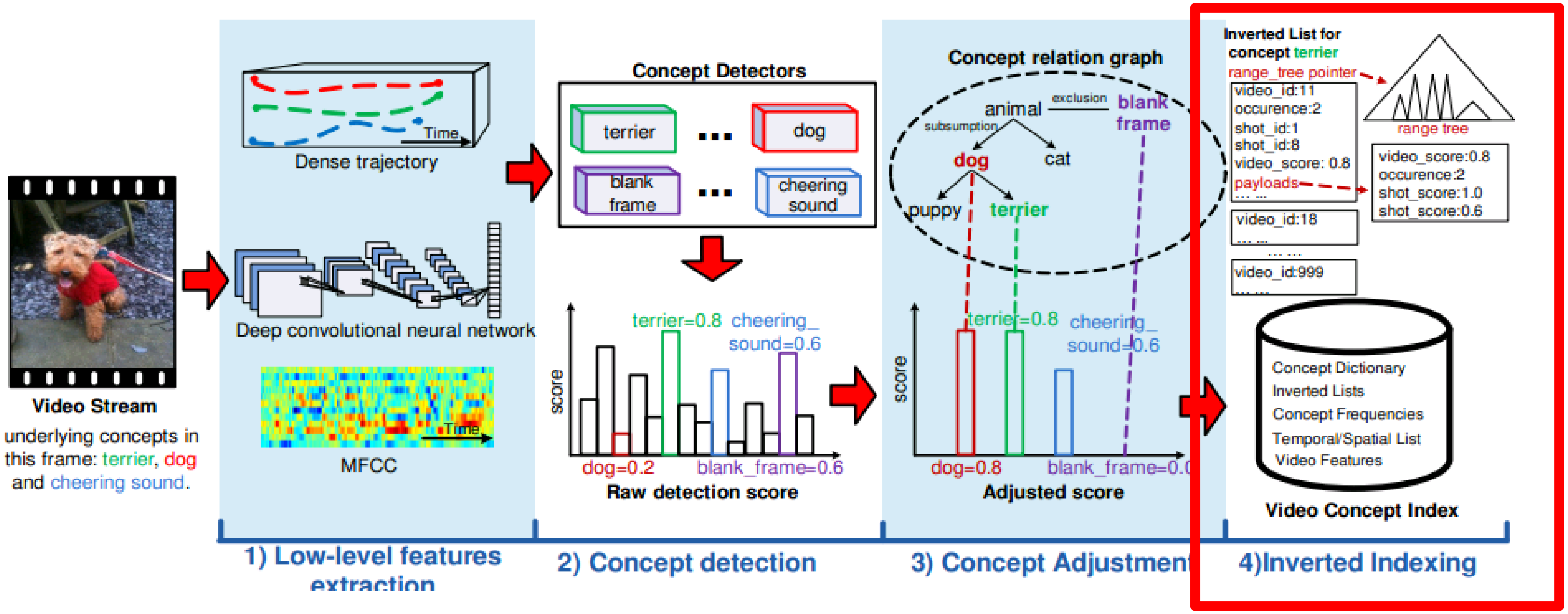
Q2: Does the process converge? If so, to where

A: Yes, to the local optimum. See the theorem i

Q3: Does the arbitrarily predefined weighting scheme converge?

A: Not guaranteed, but the discussed weights guarantees the convergence.

Indexing Semantic Features



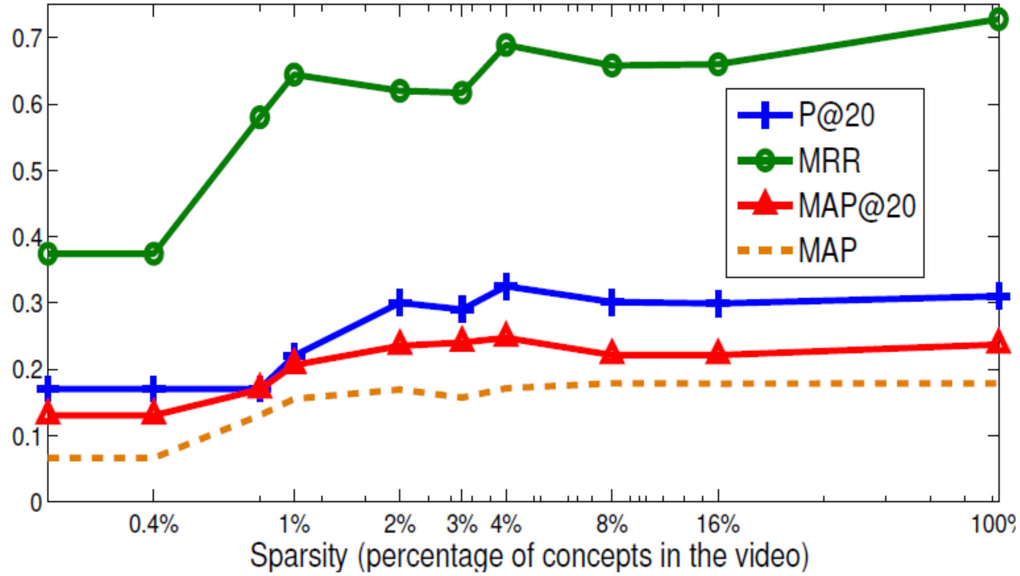
- Finally, the adjusted concept representation is indexed by the inverted index. Indexing the real-valued score. Our index supports:
  - modality search: visual:dog, ocr:dog
  - score range search: score(dog, >=, 0.7)
  - basic temporal search: tbefore(dog, cat), twindow(3s,dog, cat)
  - Boolean logical search: dog AND NOT score(cat, >=, 0.5)

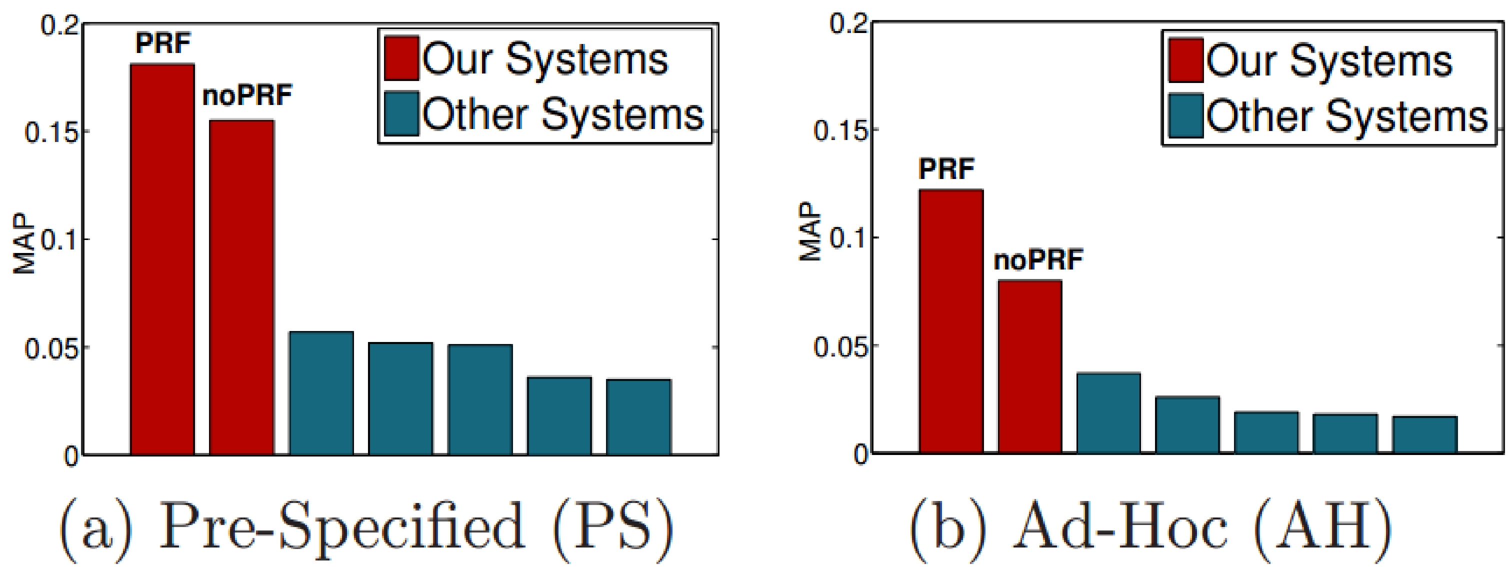
Related Work

- Categorization of reranking methods:
  - Classification-based
    - [Yan et al. 2003] [Hauptmann et al. 2008][Jiang et al. 2014]
  - Clustering-based
    - [Hsu et al. 2007]
  - LETOR(LEarning TO Rank)-based
    - [Liu et al. 2008][Tian et al. 2008][Tian et al. 2011]
  - Graph-based
    - [Hsu et al. 2007][Nie et al. 2012]

R. Yan, A. G. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *CVIR, 2003.*
A. G. Hauptmann, M. G. Christel, and R. Yan. Video retrieval based on semantic concepts. *Proceedings of the IEEE,* 96(4):602–622, 2008.
L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *ICMR, 2014*
W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. In *Multimedia, 2007.*
Y. Liu, T. Mei, X.-S. Hua, J. Tang, X. Wu, and S. Li. Learning to video search rerank via pseudo preference feedback. In *ICME, 2008.*
X. Tian, Y. Lu, L. Yang, and Q. Tian. Learning to judge image search results. In *Multimedia, 2011.*
X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *Multimedia,* 2008.
W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. In *Multimedia, 2007.*
L. Nie, S. Yan, M. Wang, R. Hong, and T.-S. Chua. Harvesting visual concepts for image search with complex queries. In *Multimedia, 2012.*

Impact of the model parameters





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

Limitations

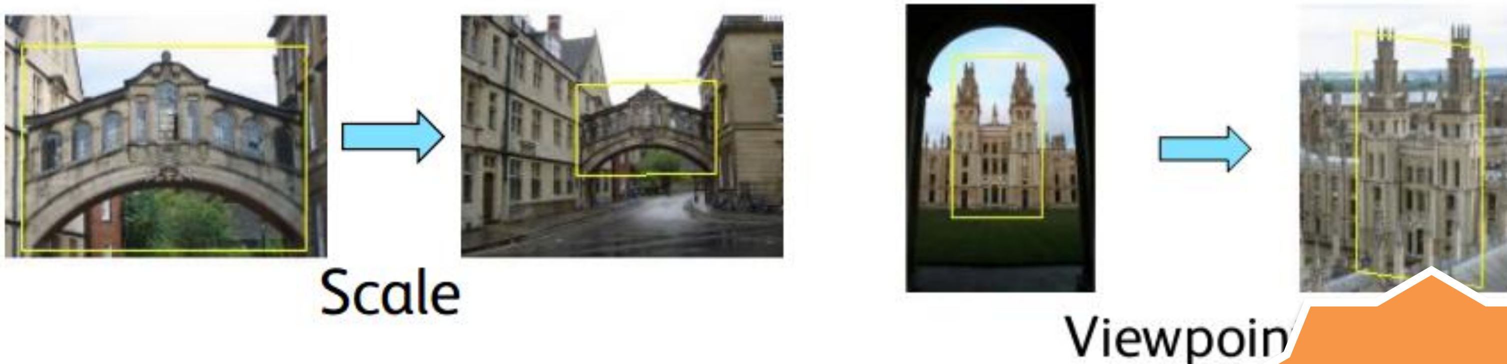
- The learning philosophy may not apply to

Related Work

- Related problems:
  - Content-based Image Retrieval
  - Copy Detection
  - Semantic Concept Indexing / Action Detection
  - Multimedia Event Detection

(Disclaimer: brief overview of related problems)

Content-based Image Retrieval



well-studied problem

- **Goal:** find visually similar images [Sivic et al 2006]
- **Query**: a single image (query-by-example)
- Single Modality. Minimum semantic understanding.
- Instance search: search the key frames about a specific instance [Zhu et al 2012]

Copy Detection/
Near Duplicate Detection

- **Goal:** find video copies derived from the input video, usually by means of transformations such as addition, deletion, formatting modification, etc [Over et al 2008].
- **Query**: a segment of video.
- Multimodal. Minimum semantic understand

well-studied problem

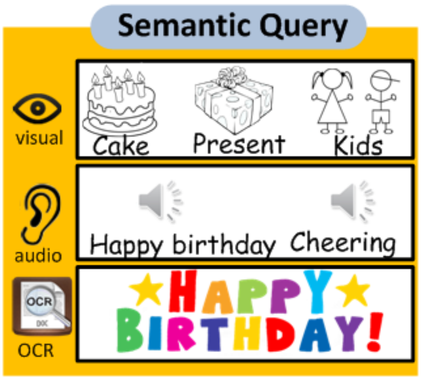
Semantic Concept Detection/ Action Detection

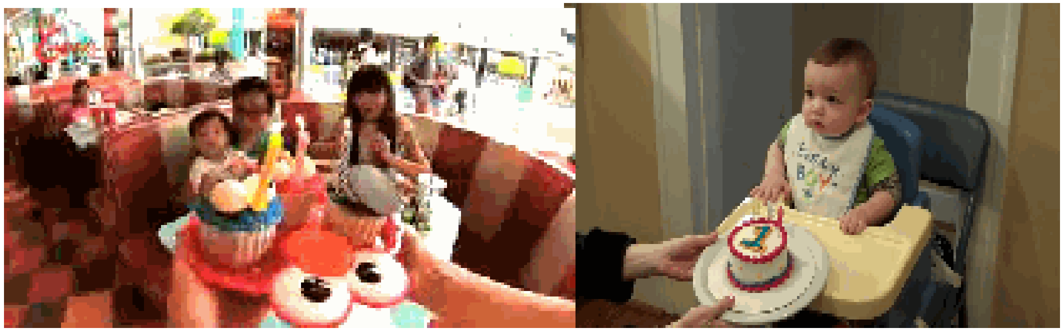
dog

- **Goal:** find segments of video that contains the concept.
- **Query**: a concept name or ID.
- Simple Query.
- The key is to build accurate individual detectors.
- Need a lot of training data.

Multimedia Event Detection (MED)

Birthday party





- **Goal:** find video about certain complex events [Over 2014]. Initiated by NIST TRECVID in 2012.
- **Query**: text or example videos about an event.
- Complex query.
- Solving the problems need semantic understanding about video content (especially for semantic queries).

Generalized MED Problem

- The proposed problem is a generalized Multimedia Event Detection (MED) problem.
- It is similar to MED but with the following differences:
  - The query can be about everything, not necessarily just an event.
  - Expand the boundary from large-scale to web-scale.