Fast and Accurate Content-based Semantic Search in 100M Internet Videos

Lu Jiang[1], Shoou-I Yu[1], Deyu Meng[2], Yi Yang[3], Teruko Mitamura[1], Alexander Hauptmann[1]

[1] **Carnegie Mellon University**

[2] **Xi'an Jiaotong University**

[3] **University of Technology Sydney**

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

- Introduction
- Proposed Approach
- Experimental Results
- Conclusions

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]

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.

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-
  ... may

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

  - 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

- We address a 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) or video examples.

Example Queries

- In response to a query, 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

Introduction

- We study a 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) or video examples.
- We are interested in searching hundreds of millions of videos within the maximum recommended waiting time for a user, i.e. 2 seconds [Nah, 2004], while maintaining maximum accuracy.

Fiona Fui-Hoon Nah. A study on tolerable waiting time: how long are web users willing to wait? Behaviour & Information Technology, 23(3):153–163, 2004.

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)

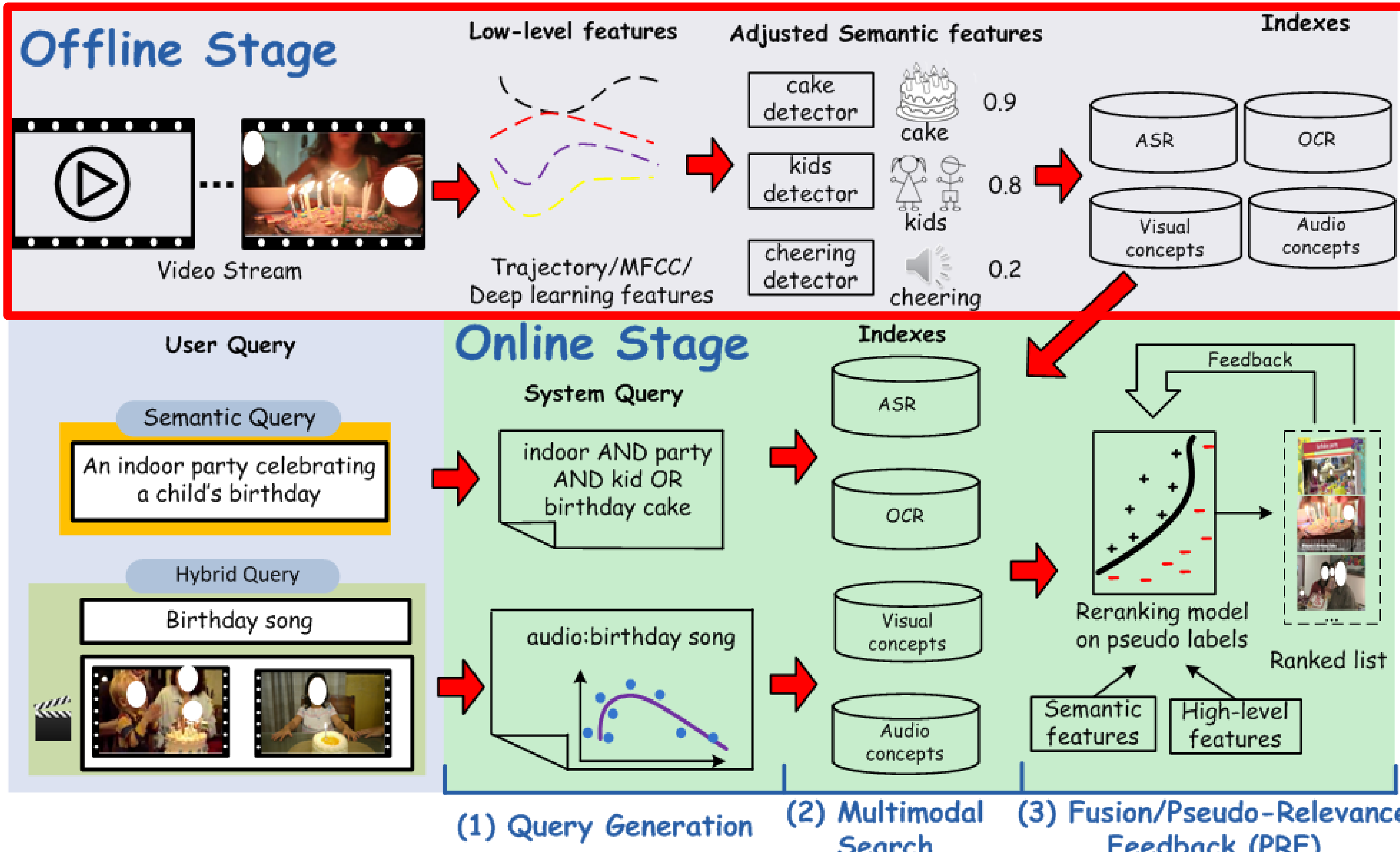
Result Overview

- We propose a novel and practical solution that can
  - Scale up the search to hundreds of millions of Internet videos.
    - 0.2 second to process a semantic query on 100 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 and 2014, the most representative task on this topic.
  - To the best of our knowledge, it is **the first content-based video retrieval system** that can search a collection of 100 million videos.

Outline

- Introduction
- **Proposed Approach**
- Experimental Results
- Conclusions

Framework



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.

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.

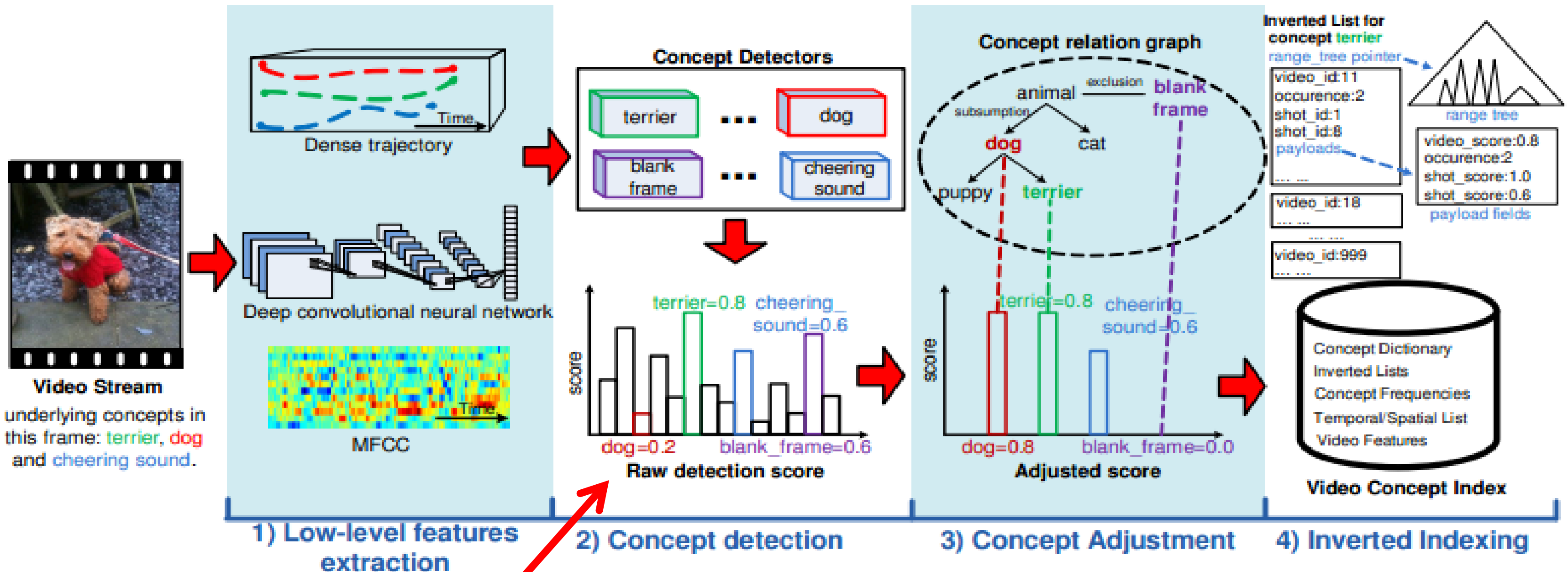
- Indexing semantic concepts is not well understood.
- Existing methods index the raw detection score of semantic concepts by dense matrices [Mazloom et al. 2014][Wu et al. 2014][Lee et al. 2014]
- We propose a scalable semantic concept indexing method. The key is a novel method called concept adjustment.

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

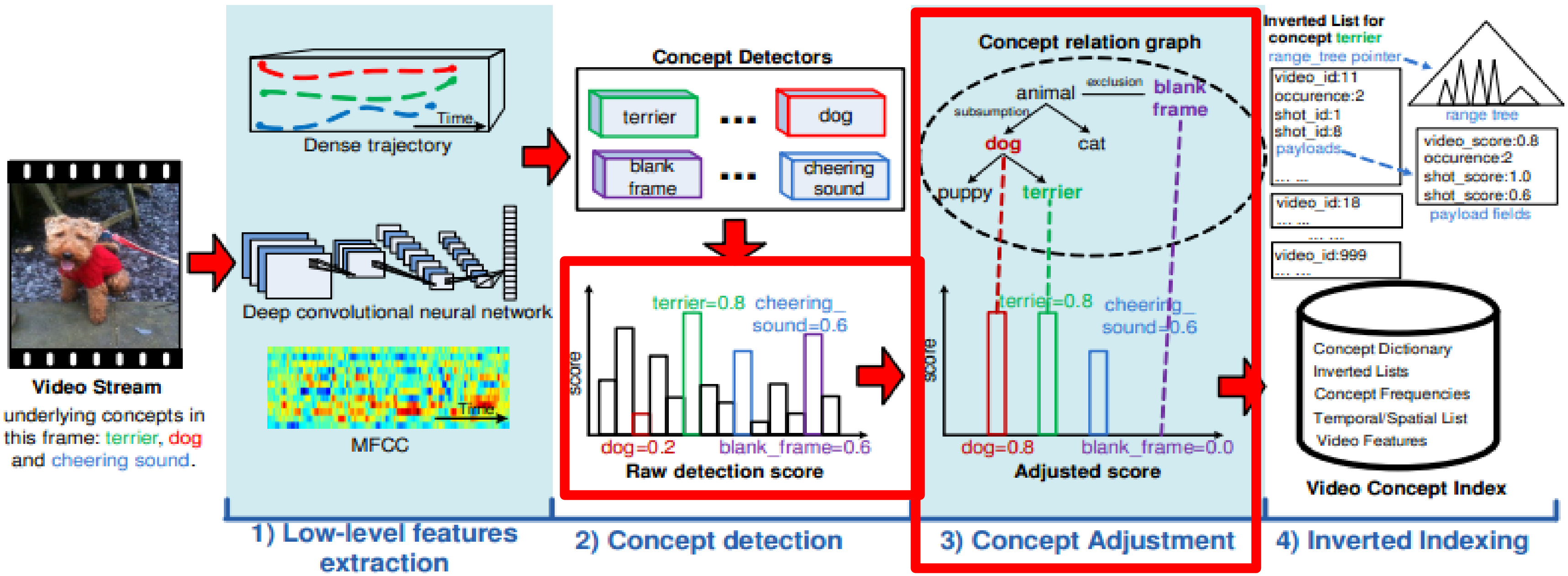


steps in the offline video indexing.

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**.
- Normalization :

Indicator function

$$\hat{v}_i = \min\left(1, \frac{v_i}{\sum_{j=1}^m v_j} \sum_{j=1}^m f_p(\mathbf{D})_j I(v_j)\right)$$

Concept Adjustment Model: Distributional Consistency

- A naive regularizer→ infeasible to solve.

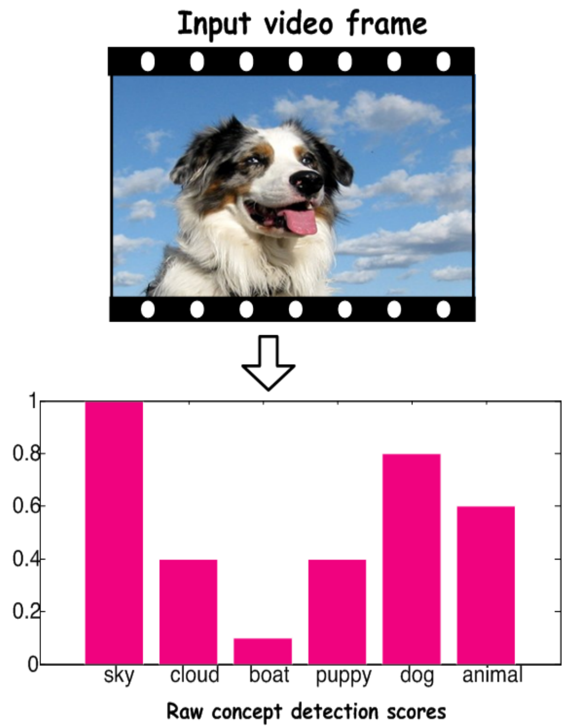
$$g(\mathbf{v};\alpha,\beta) = \frac{1}{2}\beta^2\|v\|_0$$

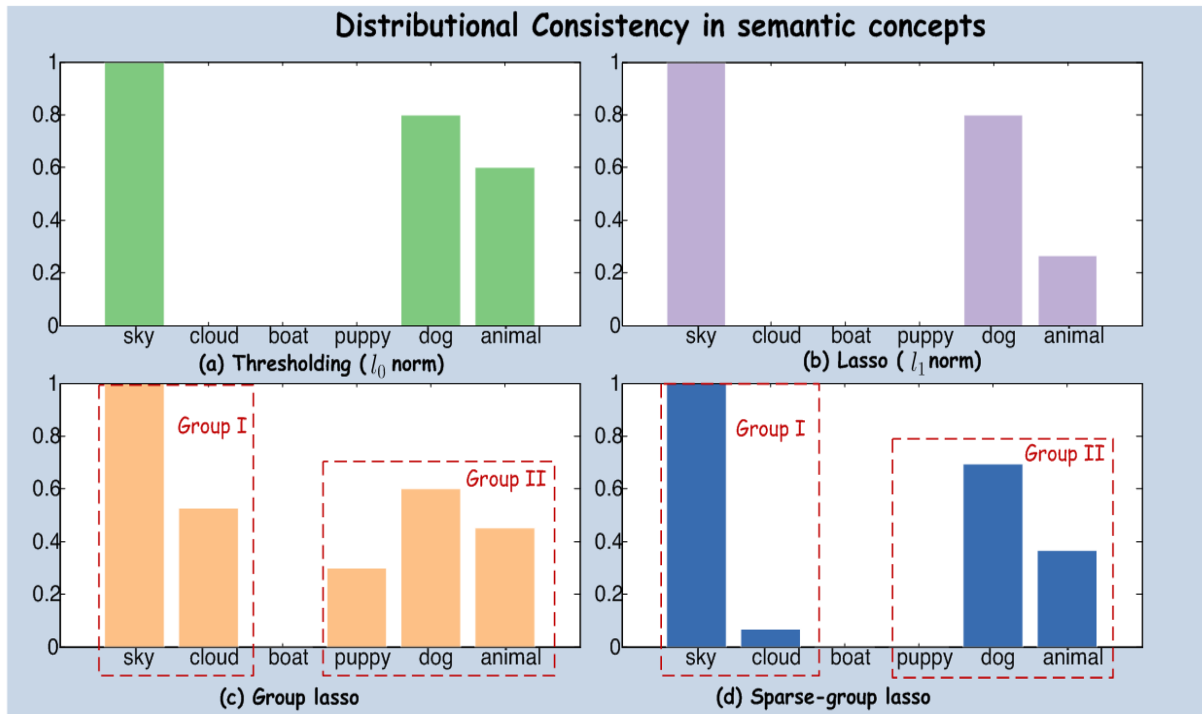
- A more general regularizer :

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

  – 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)

Distributional Consistency: A Toy Example





All the adjustment methods above special cases of our adjustment model.

Concept Adjustment Model: Distributional Consistency

- A more general regularizer :

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

  - 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.
We can cluster the concepts in their training data to get the co-occurring groups.

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

- The proposed adjustment model is:

distributional consistency

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

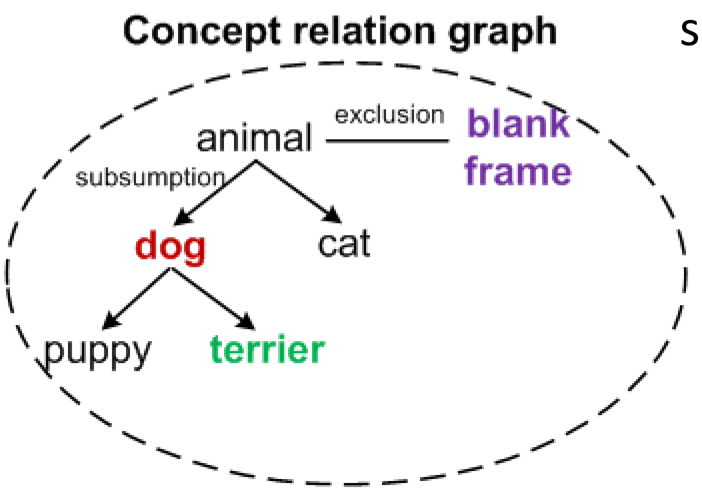
logical consistency

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: 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]



subsumption

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

exclusion **only make sense for shot-level features.**

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

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

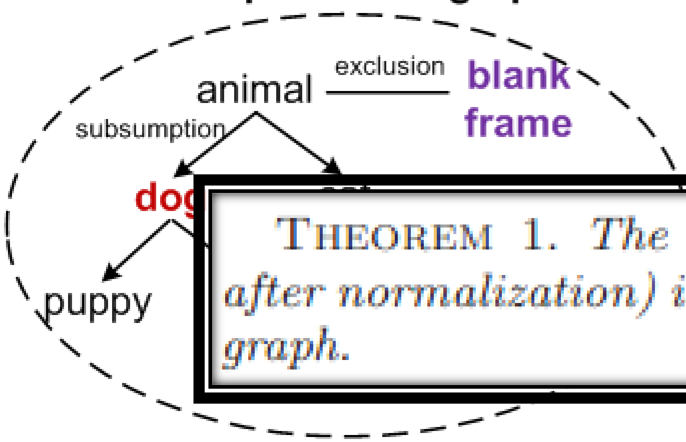
Integer programming
solved by mix-integer toolbox or by constraint relaxation.

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.*

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 1. *The optimal solutions of Eq. (1) (before or after normalization) is logically consistent with its given HEX graph.*

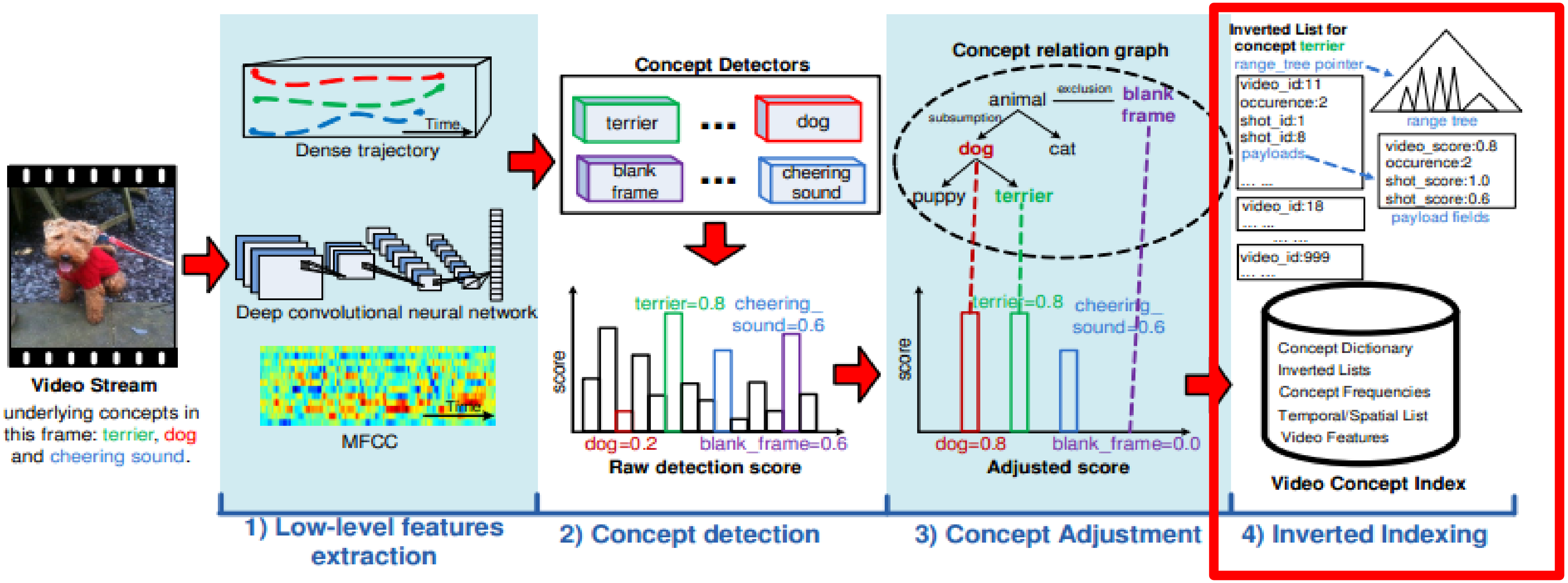
features.

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

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

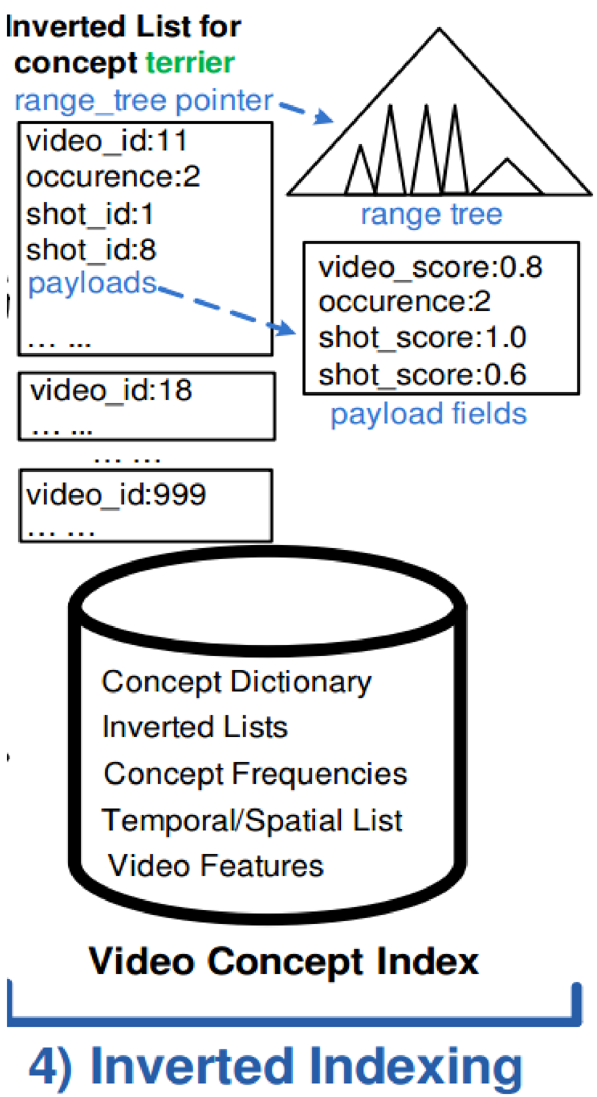
Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *ECCV, 2014.*

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.

Indexing Semantic Features



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)

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 other semantic concept features.
  - Raw prediction scores of the 3000+ concepts trained in [Jiang et al. 2015].

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

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 other
  - Raw prediction scores of the 3000+ 2015].

```
Actor implies Person
Adult implies Person
Airplane_Flying implies Airplane
Airplane implies Vehicle
```

```
Black_Frame excludes Animal
Black_Frame excludes Bridges
Black_Frame excludes Building
```

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

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 MED

Better performances

Comparison of the full adjustment model with its special case top-k thresholding

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

The MAP is low because here we only use 346 semantic features.

Experiments on the SIN dataset

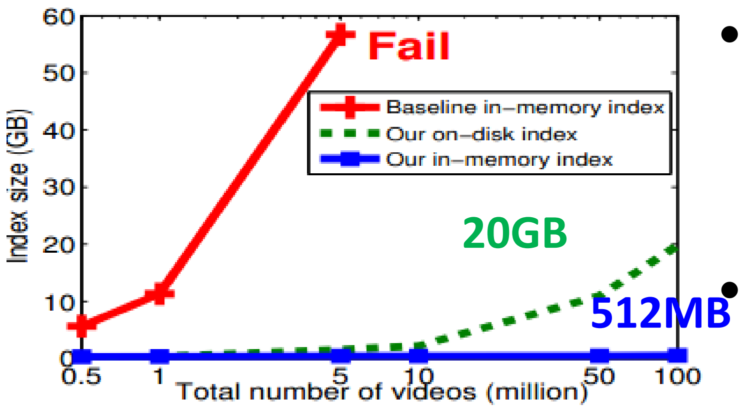
- We test adjustment method on TRECVID SIN dataset, where the ground-truth labels on each video shot are available.
- Test on 1500 shots in 961 videos. Evaluated by Root Mean Squared Error (RMSE).

Method	RMSE
Raw Score	7.671
HEX Graph Only	8.090
Thresholding	1.349
Top-k Thresholding	1.624
Group Lasso	1.570
Our method	**1.236**

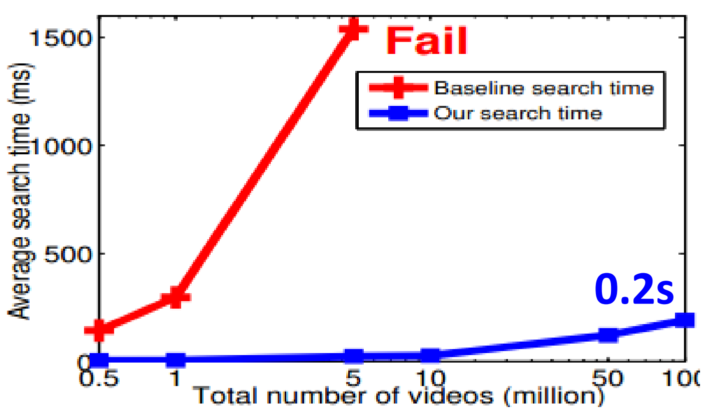
The proposed method is more accurate than the baseline methods.

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.

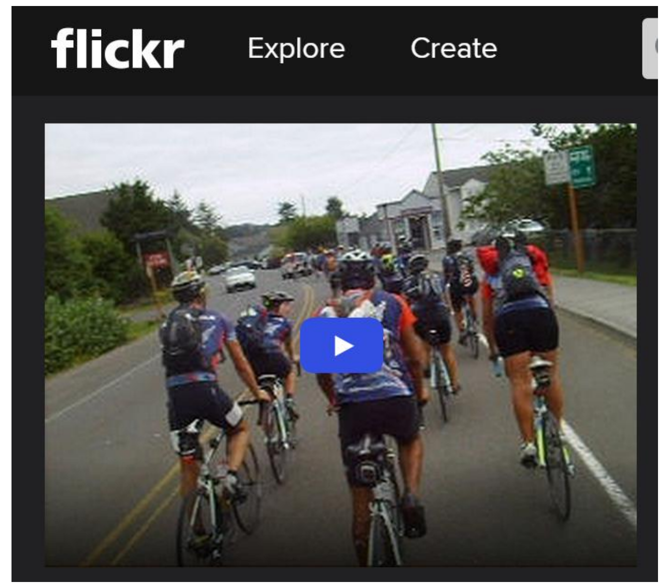
The proposed method is scalable and efficient.

Experiments on YFCC (Yahoo Flickr Creative Commons)

- We manually created queries for 30 products.
- Put commercials about the product to related video (in-video ads.)
- Search over 800K videos in the dataset.

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

Put ads in relevant videos on Flickr.



Queries and more results are available at:

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

Experiments on YFCC

- We manually created queries for 30 products.
- Put commercials about the product to related video (in-video ads.)
- Search over 800K videos in the dataset.
- Evaluate the relevance of the top 20 returned results.

Average performance for 30 commercials on YFCC

Category	#Ads	Evaluation Metric		
		P@20	MRR	MAP@20
Sports	7	0.88	1.00	0.94
Auto	2	0.85	1.00	0.95
Grocery	8	0.84	0.93	0.88
Traveling	3	0.96	1.00	0.96
Miscellaneous	10	0.65	0.85	0.74
Average	30	0.81	0.93	0.86

Queries and more results are available at:

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

Experiments on YFCC

Queries and more results are available at:

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

Outline

- Introduction
- Proposed Approach
- Experimental Results
- Conclusions

Conclusions

- 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.
- Take home: experimental results show that our system can search 100 million Internet videos within 0.2 second.
- We share our concept features of the 0.8 million videos in the YFCC dataset.

Features:

*Please cite the corresponding papers for using our features (800,000 Internet videos in YFCC100M).

Concept Features	Raw	Adjusted
YFCC100M (609 concepts) [1,3,4]	features, dictionary for all semantic concepts	features
Google Sports (478 concepts) [1,3,5]	features, dictionary for all semantic concepts	features
IACC (346 concepts) [1,3,6]	features, dictionary for all semantic concepts	features
DIY (1601 concepts) [1,3,7]	features, dictionary for all semantic concepts	features

THANK YOU.
QUESTIONS?