Zero-example Video Search

Chong-Wah Ngo
City University of Hong Kong

James Yi-Jie Lu, Maaike de Boer and Hao Zhang
Agenda

• Problem definition
• Quick overview of problems, methods, results
• Baseline to be shared

The talk will be delivered in ICMR 2017 as an 1-hour tutorial, along with few other talks: 1) Semantic Indexing, 2) Ad-hoc search, 2) Instance Search, 3) Multimedia Event Detection, 4) Video-to-Text, with a number of baselines to be released. The series of talks is to encourage more TRECvid participation, by providing tools/open sources to rapidly setup a decent system.
Problem

Given a textual query, find the relevant video clips from large video collection.

**Query:** Find shots of something **burning** with **flames** visible

explosion?

smoke?
Computers are not as *smart* as humans…

- Between computable low-level features and high-level semantics

Scalability (speed)

- Thousands of semantic categories
- Billions of images/videos on the web
Semantic Gap

User queries

Model

Model?

Pre-train

Low-level visual features

Extract

Raw images/snippets
Concept-based video search

• Database indexing

• Online search

Find “people talking”

Concept Bank

Anchor person

Person, Meeting, …

Military action, Vehicle, Road, Building…
The idea of zero-example video search starts from TRECVID

**TRECVID**
- Sponsored by NIST, USA
- Provide benchmark and evaluation annually for system evaluation

**TRECVID dataset**
- Broadcast News (NTV, CCTV, MSNBC, CNN…) 2003-2006
- Documentary Videos from the Netherlands. 2007-2008
- Web Videos: 2010 and beyond
# TREC Video Retrieval Evaluation (TREVID)

<table>
<thead>
<tr>
<th>TV</th>
<th>Data domain</th>
<th>Devel. set (keyframe #)</th>
<th>Test set (keyframe #)</th>
<th># of evaluated queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>05</td>
<td>Broadcast news</td>
<td>80h (43,873)</td>
<td>80h (45,765)</td>
<td>24</td>
</tr>
<tr>
<td>06</td>
<td>Broadcast news</td>
<td>--</td>
<td>80h (79,484)</td>
<td>24</td>
</tr>
<tr>
<td>07</td>
<td>Documentary</td>
<td>50h (21,532)</td>
<td>50h (18,142)</td>
<td>24</td>
</tr>
<tr>
<td>08</td>
<td>Documentary</td>
<td>--</td>
<td>100h (35,766)</td>
<td>48</td>
</tr>
<tr>
<td>16</td>
<td>Internet archive (AVS)</td>
<td>1,400h</td>
<td>600h (335,944)</td>
<td>30</td>
</tr>
<tr>
<td>13-16</td>
<td>User-generated (MED)</td>
<td>416h</td>
<td>849h (MED14Test)</td>
<td>20 (predefined) + 10 (ad-hoc)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7,580h (full)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1,244h (sub)</td>
<td></td>
</tr>
</tbody>
</table>
Ad-hoc Query

Event and/or Person-Things

Find shots with a person walking or riding a bicycle

Selected Concepts

- Person, Bicycle, Individual, Walking, Running, Backpacker...
- Walking, Walking_Running, Person, Bicycle, Car, Group, Motorcycle...
- Bicycles, Person, Walking, Walking_Running, Horse, Dog...
- Bicycles, Person, Walking, Walking_Running, Daytime_outdoor...

Return top-1,000 shots

Wordnet  Google search  Flickr context  transfer learning
Multimedia Event Query

Complex and generic events occurring at a specific place and time involving people interacting with other people / objects.
High Variability

Changing a vehicle tire

Marriage proposal
Changing A Vehicle Tire

- **Description**
  One or more people work to replace a tire on a vehicle

- **Explication**
  The process for replacing a tire includes removing the existing tire and installing the new tire onto the wheel of the vehicle ......

- **Evidential description**
  - **Scene**: garage, outdoors, street, parking lot
  - **Objects/people**: tire, lug wrench, hubcap, vehicle, tire jack
  - **Activities**: removing hubcap, turning lug wrench, unscrewing bolts
  - **Audio**: sounds of tools being used; street/traffic noise
Changing A Vehicle Tire

- **Event**: Changing a vehicle tire
  - **Interaction**
    - **Human**
    - **Object**
    - **Scene**
      - **Low-level motion features**
      - **Low-level visual features**
    - **Action**
      - Squatting
      - Standing up
      - Walking
    - **Scene**
      - Side of the road
      - Tire wrench
      - Tire
  - **Event**
    - Changing a vehicle tire
**Issue-A: Concept Bank**

1. How to determine the list of concepts to index?
2. How large should the concept bank be?
3. What are the expected classification accuracy?
How many concepts required for “Happy Birthday”? 

- Birthday cake
- Gift
- Decoration: Balloon
- Decoration: Party hat
- Several persons gathered around
- Candle
- Birthday cake
- Gift
General guidelines

- A mixed of general and specific concepts
- A mixed of concepts with different complexities (from event-oriented concepts to objects and scenes)
- General versus specific
  - General concepts are more important than specific concepts
- Quantity versus quality
  - Include more concepts than to improve the quality of individual concepts
- Summary: A large and diverse concept bank
**Issue-A: Concept Bank**

- **Lower level**
  - ImageNet 1000: objects, activities
  - SIN 346 (AVS): objects, activities

- **Mid level**
  - FCVID 239: activities, events
  - Places 205: scenes

- **Higher level**
  - Sports 487: activities, events
  - RC 497 (MED): mixed
Number of concepts vs. MAP

MED14Test (oracle run of top-$k$ concepts)
**Number of concepts vs. MAP**

**MED14Test** (oracle run of top-\(k\) concepts)
The performance of high level concepts increases sharper than mid and low level concepts.
**Issue-B: Concept Selection**

1. How to select the most appropriate concept detectors?
2. How many concepts are enough?
3. How to combine concepts?

Find shots of something **burning** with **flames** visible

- explosion?
- smoke?
- fire?
- screaming?
Issue-B: Concept Selection

Semantic Query Generation

- Concept matching and selection

NLP Parser

Tokens

Concept Bank

- max or average pooling

Video Corpus

Concept-based Video Representation

- Cleaning
- Spray bottle
- Refrigerator
- Dishwasher
- Stove
- Microwave
- Kitchen

Event Query

Cleaning an appliance

Online

Offline
B1: Semantic Query Generation

Reasoning
- Exact Match
- Ontology
- Synonym
- Word specificity

Machine learning
- Word2vec
- VideoStory

Semantic Query

< Objects >
- Bike 0.60
- Motorcycle 0.60
- Mountain bike 0.60

< Actions >
- Bike trick 1.00
- Ridding bike 0.62
- Flipping bike 0.61
- Assembling a bike 0.60

< Scenes >
- Motorcycle speedway 0.01
- Parking lot 0.01
- ......
## B1: Ad-hoc Query

<table>
<thead>
<tr>
<th>Query</th>
<th>WordNet (WUP)</th>
<th>Google Search (NGD)</th>
<th>Flickr Context (FCS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A goal in a soccer match</td>
<td>Striking</td>
<td>Sports</td>
<td>Soccer</td>
</tr>
<tr>
<td>Something burning with flames visible</td>
<td>Sky</td>
<td>Soldiers</td>
<td>Smoke</td>
</tr>
<tr>
<td>Scenes with snow</td>
<td>Landscape</td>
<td>Person</td>
<td>Urban_Scenes</td>
</tr>
<tr>
<td>A train in motion</td>
<td>Vehicle</td>
<td>Car</td>
<td>Railroad</td>
</tr>
</tbody>
</table>

Query #199: Find shots with a person walking or riding a bicycle

Selected Concepts:

Bicycles (0.31), Person (0.31), Walking (0.12), Walking_Running (0.10), Horse (0.06), Dog (0.05)...

Top 10 ranked shots
AP=0.323
B1: Event Query

Cleaning an appliance
B1: Concept Selection

“How Hiking”

How to convert “context” into a searchable query?

**B1: Concept Selection**

- Exact or partial string matching

(hope for better precision)

- Ontology reasoning
  - Wordnet, Conceptnet
- Context inference
  - Word2Vec, Flickr, Wikipedia

(hope for better recall)

“dog show” to “dog show” (1.0)
“dog show” to “dog” (0.5)

Context understanding is difficult, and can easily cause query drift when context is interpreted wrongly.
B1: Why Wordnet reasoning is risky?

Dog Show

Concept "dog"

red wolf
kit fox
cat
horse
mammal
carnivore
animal

ImageNet
SIN
B1: Why Conceptnet is risky?
B2: How many concepts are enough?

Event 31: Beekeeping
B2: How many concepts are enough?

Event 31: Beekeeping

The diagram shows the average precision for different top $k$ concepts. The precision increases as more concepts are considered, reaching a peak around $k=8$ before decreasing as more concepts are added. The event is labeled as Beekeeping.
B2: How many concepts are enough?

Approaches

– Thresholding – simple fix
– Manual screening
– Evidential pooling
– Incremental Word2Vec
Event 31: Beekeeping
## B2: A simple fix – Naïve cutoff

<table>
<thead>
<tr>
<th>Concept Bank</th>
<th>#Concepts</th>
<th>Optimum k</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>487</td>
<td>10</td>
<td>0.103</td>
</tr>
<tr>
<td>FCVID</td>
<td>239</td>
<td>1</td>
<td>0.071</td>
</tr>
<tr>
<td>Research Collection</td>
<td>497</td>
<td>2</td>
<td>0.053</td>
</tr>
<tr>
<td>ImageNet</td>
<td>1,000</td>
<td>3</td>
<td>0.049</td>
</tr>
<tr>
<td>Places</td>
<td>205</td>
<td>2</td>
<td>0.020</td>
</tr>
<tr>
<td>SIN</td>
<td>346</td>
<td>5</td>
<td>0.014</td>
</tr>
<tr>
<td>Concept Bank (ALL)</td>
<td>2,774</td>
<td>9</td>
<td>0.129</td>
</tr>
<tr>
<td>AutoSQGSys [1]</td>
<td>4,043</td>
<td>--</td>
<td>0.115</td>
</tr>
</tbody>
</table>

**The MAP is reported on MED14Test**

B2: Manual concept screening

- Remove *false positives* by screening the names of concepts
- Only include concepts that are *distinctive* to an event if we find a concept detector *semantically matches* the event
- Remove concepts for which training videos may appear in *very different context* based on human’s common sense

**Rock climbing**

- Rock climbing, bouldering, sport climbing, artificial rock wall  
  - Relevant
- Rope climbing, climbing, rock  
  - Non-distinctive
- Rock fishing, rock band performance  
  - False positive
- Stone wall, grabbing rock  
  - Different context
Manual search is particularly useful!

Auto versus manual search

MED14Test

<table>
<thead>
<tr>
<th>MAP</th>
<th>0%</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21%</td>
<td></td>
</tr>
</tbody>
</table>
Auto versus Manual search

AVS 2016

<table>
<thead>
<tr>
<th>Mean InfAP</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
<td>1%</td>
<td>2%</td>
<td>3%</td>
<td>4%</td>
<td>5%</td>
<td>6%</td>
<td>7%</td>
</tr>
<tr>
<td>9%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.1%</td>
</tr>
<tr>
<td>8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
B2: Evidential pooling

The *evidences* that justifies a detection result only resides in a few shots

An example video of “attempting a bike trick”
If we uniformly pool a video

- We would collect the responses of many contextually relevant but indiscriminative concepts
- Lane/road, bike, riding bike, parking meter, traffic sign, guard rail, tree, cars, and pedestrians
If we pool only the query-related evidential shots

- The video representation is more focused (refined)
- Lane/road, tree, bike, riding bike, platform

B2: Evidential Pooling (Algorithm)

Index every (key)frame of a video

1. Index every (key)frame of a video

2. Event Query
   Cleaning an appliance

3. NLP Parser

4. Concept Bank

5. Concept matching and selection

6. Semantic Query Generation

7. Tokens

8. Full Semantic Query

9. Evidence Localization and Pooling

10. Concept-based Frame Representation

11. Concept-based Video Representation

12. Video Ranking
B2: Evidential Pooling (Algorithm)

Determine the most confident $k_e$ concepts

1. Event Query: Cleaning an appliance
2. NLP Parser
3. Semantic Query Generation
4. Tokens
5. Concept matching and selection
6. Concept Bank
7. Full Semantic Query
8. Select the top $k_e$ concepts
9. Evidence Localization and Pooling
10. Concept-based Frame Representation
11. Concept-based Video Representation
12. Video Corpus
13. Video Ranking
B2: Evidential Pooling (Algorithm)

Identify top-$n$ (key) frames with high responses to $k_e$ concepts.
**B2: Evidential Pooling (Algorithm)**

Perform evidential pooling to generate video-level features.
B2: Evidential Pooling (Algorithm)

Perform video search using full semantic query
### Sensitivity of $k_e$

<table>
<thead>
<tr>
<th>$k_e$</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.0725</td>
<td>0.0771</td>
<td>0.0775</td>
<td>0.0770</td>
</tr>
</tbody>
</table>

**The MAPs are based on 10 evidential shots**

The user studies show that one can recognize a complex query **with very small number of shots** (1-3).

B2: Evidential Pooling

AP of "beekeeping"

Max Pooling vs Evid Pooling

- Max Pooling
- Evidential Pooling

Noisy concepts

34 concepts total
B2: Evidential Pooling

MAP (Top k concepts)

- Evid3 (Manual)
- Evid10 (Manual)
- Evid3 (Auto)
- Evid10 (Auto)
- Max pooling
- Avg pooling
B2: Incremental Word2Vec

• Key idea
  – Vector composition using element-wise addition
  – Care about query drift more than concept similarity

  felling tree ≈ fruit tree pruning + tree frog + tree farm

  parking a vehicle ≈ vehicle + parking lot

B2: Incremental Word2Vec (Algorithm)

1. Embed a query with Word2Vec, Q
   \[ UQ = \text{Word2Vec(changing)} + \text{Word2Vec(vehicle)} + \text{Word2Vec(tire)} \]

2. Embed concepts to Word2Vec

3. Pick the most similar concept to query
   \[ C = \text{Word2Vec(most-similar-concept)} \]

4. Pick the next most similar concept
   \[ C' = C + \text{Word2Vec(next-most-similar-concept)} \]
   \[ \text{if } \text{Cosine}(UQ,C') > \text{Cosine}(UQ,C) \]
   \[ C = C' \]

Get rid of concepts which may cause query drift
### MED14Test

<table>
<thead>
<tr>
<th>Cutoff at top-X%</th>
<th>MAP</th>
<th>Ave. # of concepts</th>
<th>Stdv</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>0.142</td>
<td>1.9</td>
<td>1.3</td>
</tr>
<tr>
<td>85%</td>
<td>0.142</td>
<td>2.3</td>
<td>2.4</td>
</tr>
<tr>
<td>80%</td>
<td>0.142</td>
<td>3.0</td>
<td>4.6</td>
</tr>
<tr>
<td>75%</td>
<td>0.141</td>
<td>3.8</td>
<td>6.3</td>
</tr>
<tr>
<td>50%</td>
<td>0.137</td>
<td>7.2</td>
<td>12.1</td>
</tr>
<tr>
<td>25%</td>
<td>0.136</td>
<td>9.3</td>
<td>13.4</td>
</tr>
<tr>
<td>0%</td>
<td>0.136</td>
<td>9.4</td>
<td>13.4</td>
</tr>
</tbody>
</table>

MAPs are quite stable over different cutoff points
**B3: How to combine concepts?**

Weighted sum of concept scores

$$W_1 \times \text{Railroad} + W_2 \times \text{Bridges} + W_3 \times \text{Tunnel} = \text{Train}$$

**Open issue**

- How to normalize the score of concepts from different datasets and trained using different classification methods?
- How to combine concepts of similar names, probably learnt with different training examples of different context, from different datasets?
B3: How to combine concepts?

• AND-OR (Wasade@TRECVid2016)
  – OR: max operator
  – AND: sum or multiplying operator

Query: One or more people walking or bicycling on a bridge during daytime

people and (walking or bicycling) and bridge and daytime
**OEx Baseline**

- For both ad-hoc and multimedia event queries
- Support more than 10,000 visual concepts
- Highly efficient: Can search a complex query within seconds on a laptop
- Support *concept screening* and *interactive search*
- Publicly available
  - Open source
  - Console for interactive search
  - Concept features for MED14Test, IACC, TV08

Baseline: Concept Bank

- Places 205
- SIN 346
- RC 497
- ImageNet 1,000
- ImageNet 12,998
- FCVID 239
- Sport 487

- **ResNet-50**
- **GoogLeNet** (ImageNet-Shuffle 13K)
- **FC7 (AlexNet) + SVM**
- **3D CNN**
Baseline: Concept Selection

Initialization

Create query-concept terms table

Automatic concept matching

Concept screening?

Hand-pick concepts

Video ranking

End

Max pooling of keyframe scores
Baseline: Concept Selection

1. Each token votes for its most relevant concepts by a score metric concerning TF-IDF, word specificity, WordNet synonyms, etc. [1]

Baseline: Concept Selection

1. Each token votes for its most relevant concepts by a score metric concerning TF-IDF, word specificity, WordNet synonyms, etc. [1]

Baseline: Concept Selection

Query Tokens

- bike
- motorized bike
- motorcycle
- bike jump
- street
- parking lot
- concrete floor
- bike ramps
- riding
- one wheel standing
- jumping
- spinning
- flipping

Concept Names

- bicycle
- motorcycle
- bike trick
- street
- riding bike
- freestyle bmx bike

1. Each token votes for its most relevant concepts by a score metric concerning TF-IDF, word specificity, WordNet synonyms, etc. [1]

## Baseline: Concept Weighting

### Query Tokens
- bike
- motorized bike
- motorcycle
- bike jump
- street
- parking lot
- concrete floor
- bike ramps
- riding
- one wheel standing
- jumping
- spinning
- flipping

### Concept Names
- bike trick
- riding bike
- freestyle bmx bike
- motorcycle racing
- bicycle
- motorcycle
- street
- .......
- .......
- .......
- .......
- .......
- .......

2. Weight the selected concepts according to the votes. [1]

---

Baseline: Console

4s /query for a concept bank (14K), including query processing, searching, ranking
**AVS 2016 Concept Bank**

### Basic
- Places 205
- SIN 346
- Research collection 497
- ImageNet 1,000

\[2,048 \text{ in total}\]

### Large
- Places 205
- SIN 346
- Research collection 497
- ImageNet 12,998

\[14,046 \text{ in total}\]
AVS 2016 Performance

On TRECVID-2016 AVS Benchmark

- **Manual run with AND-OR operators**
- **Fully-automatic run on the large (15K) concept bank**
- **Manually-assisted run on the basic (2K) concept bank**
- **Fully-automatic run on the basic (2K) concept bank**

**Scores of the other teams are from the official report of TRECVID-2016 AVS task. The blue bars are manually-assisted runs and orange bars are fully-automatic runs.**
On TRECVID-2008 Video Search Task Benchmark (Fully-Automatic)

** Scores of the other teams are from the official report of TRECVID-2008 search task.
MED14 Test Concept Bank

**Basic**
- Places 205
- SIN 346
- Research collection 497
- ImageNet 1,000

2,048 in total

**Large**
- Places 205
- SIN 346
- Research collection 497
- ImageNet 1,000
- FCVID 239
- Sport 487

2,774 in total
MED14Test Performance

On MED14Test 0-Ex Benchmark

<table>
<thead>
<tr>
<th>Concept Bank</th>
<th>#Concepts</th>
<th>Query</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (basic)</td>
<td>2,048</td>
<td>Automatic</td>
<td>0.069</td>
</tr>
<tr>
<td>Baseline (basic)</td>
<td>2,048</td>
<td>Manual</td>
<td>0.110</td>
</tr>
<tr>
<td>Baseline (large)</td>
<td>2,774</td>
<td>Automatic</td>
<td>0.113</td>
</tr>
<tr>
<td>Baseline (large)</td>
<td>2,774</td>
<td>Manual</td>
<td>0.191</td>
</tr>
<tr>
<td>AutoSQGSys [1]</td>
<td>4,043</td>
<td>Automatic</td>
<td>0.115</td>
</tr>
<tr>
<td>Dynamic composition [2]</td>
<td>3,135</td>
<td>Automatic</td>
<td>0.134</td>
</tr>
<tr>
<td>Incremental Word2Vec [3]</td>
<td>2,277</td>
<td>Automatic</td>
<td>0.142</td>
</tr>
<tr>
<td>VisualSys [1]</td>
<td>4,043</td>
<td>Manual</td>
<td>0.176</td>
</tr>
</tbody>
</table>


Conclusion

What do we gain over the past one decade?

- Bigger dataset
- Bigger concept bank (more than 5,000 concepts!)
- Better and more variety of features
- Still a large room for performance improvement
  (MAP hardly > 0.2)
- Still, simple approach works better
- Some light on number of concepts to select for a query
- Context is still difficult

Try the baseline 😊😊😊😊

http://vireo.cs.cityu.edu.hk/zeroex/ or
https://github.com/iiedii/0-ex