Zero-Example Event Detection and Recounting

Speaker: Yi-Jie Lu

Yi-Jie Lu, Hao Zhang, Ting Yao, Chong-Wah Ngo
On behalf of VIREO Group, City University of Hong Kong
Feb. 12, 2015
Outline

- Multimedia Event Detection (MED)
  - Background
  - System Overview
  - Findings

- Multimedia Event Recounting (MER)
  - Background
  - System Workflow
  - Results
Background

- A Multimedia Event
  - An activity occurring at a specific *place* and *time* involving *people* interacting with other *people* / *objects*.

<table>
<thead>
<tr>
<th>Pre-Specified Testing and Evaluation Events</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PS12: E021-E030</strong></td>
</tr>
<tr>
<td>E021 - Bike trick</td>
</tr>
<tr>
<td>E022 - Cleaning an appliance</td>
</tr>
<tr>
<td>E023 - Dog show</td>
</tr>
<tr>
<td>E024 - Giving directions</td>
</tr>
<tr>
<td>E025 - Marriage proposal</td>
</tr>
<tr>
<td>E026 - Renovating a home</td>
</tr>
<tr>
<td>E027 - Rock climbing</td>
</tr>
<tr>
<td>E028 - Town hall meeting</td>
</tr>
<tr>
<td>E029 - Winning race without a vehicle</td>
</tr>
<tr>
<td>E030 - Working on a metal crafts project</td>
</tr>
</tbody>
</table>
• A Multimedia Event
  – An activity occurring at a specific *place* and *time* involving *people* interacting with other *people / objects*.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E021 - Bike trick</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>E022 - Cleaning an appliance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E023 - Dog show</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E024 - Giving directions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E025 - Marriage proposal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E026 - Renovating a home</td>
<td></td>
<td><strong>E036 - Felling a tree</strong></td>
</tr>
<tr>
<td>E027 - Rock climbing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E028 - Town hall meeting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E029 - Winning race without a vehicle</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>E030 - Working on a metal crafts project</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>E031 - Beekeeping</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E032 – Wedding shower</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E033 – Non-motorized vehicle repair</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E034 – Fixing a musical instrument</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E035 – Horse riding competition</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>E036 - Felling a tree</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E037 – Parking a vehicle</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E038 – Playing fetch</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E039 – Tailgating</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E040 – Tuning a musical instrument</td>
<td></td>
</tr>
</tbody>
</table>
A Multimedia Event

- An activity occurring at a specific *place* and *time* involving *people* interacting with other *people* / *objects*.

<table>
<thead>
<tr>
<th>Pre-Specified Testing and Evaluation Events</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PS12: E021-E030</strong></td>
</tr>
<tr>
<td>E021 - Bike trick</td>
</tr>
<tr>
<td>E022 - Cleaning an appliance</td>
</tr>
<tr>
<td>E023 - Dog show</td>
</tr>
<tr>
<td>E024 - Giving directions</td>
</tr>
<tr>
<td>E025 - Marriage proposal</td>
</tr>
<tr>
<td>E026 - Renovating a home</td>
</tr>
<tr>
<td>E027 - Rock climbing</td>
</tr>
<tr>
<td><strong>E028 - Town hall meeting</strong></td>
</tr>
<tr>
<td>E029 - Winning race without a vehicle</td>
</tr>
<tr>
<td>E030 - Working on a metal crafts project</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
## Background

- **A Multimedia Event**
  - An activity occurring at a specific *place* and *time* involving *people* interacting with other *people / objects*.

### Ad-Hoc Testing and Evaluation Events

<table>
<thead>
<tr>
<th>PS1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AH14: E041-E050</strong></td>
</tr>
<tr>
<td><strong>E041</strong> - Baby shower</td>
</tr>
<tr>
<td><strong>E042</strong> - Building a fire</td>
</tr>
<tr>
<td><strong>E043</strong> - Busking</td>
</tr>
<tr>
<td><strong>E044</strong> - Decorating for a celebration</td>
</tr>
<tr>
<td><strong>E045</strong> - Extinguishing a fire</td>
</tr>
<tr>
<td><strong>E046</strong> - Making a purchase</td>
</tr>
<tr>
<td><strong>E047</strong> - Modeling</td>
</tr>
<tr>
<td><strong>E048</strong> - Doing a magic trick</td>
</tr>
<tr>
<td><strong>E049</strong> - Putting on additional apparel</td>
</tr>
<tr>
<td><strong>E050</strong> - Teaching dance choreography</td>
</tr>
</tbody>
</table>
Background

- Shots for typical events
How to detect these events?
High-level Events

Model?

Low-level visual features

Extract

Raw images / video snippets
In view of concepts
In view of concepts

Several persons gathered around

Decoration: Balloon
Decoration: Party hat
Candles
Birthday cake
Gift
Gift
Views of an event

- **Event**: Changing a vehicle tire

- **Interaction**: Person opening the car trunk, Person jacking the car, Person using wrench, Person changing for a new tire

- **Action**: Squatting, Standing up, Walking

- **Object**: Tire wrench, Tire

- **Scene**: Side of the road

- **Low-level visual features**

- **Low-level motion features**

---

**VIREO Video Retrieval Group**
Zero-Example MED System
**Query Example – Changing a vehicle tire**

- [ Exemplar videos ...... ]
- Description: One or more people work to replace a tire on a vehicle
- Explication: ...
- 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
Semantic Query Generation (SQG)

- Given an event query, SQG translates the query description into a representation of semantic concepts.

Event Query:

- Attempting a Bike Trick

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

< Scenes >
- Parking lot 0.01
• Concept Bank
  – Research collection (497 concepts)
  – ImageNet ILSVRC’12 (1000 concepts)
  – SIN’14 (346 concepts)
SQG Highlights

- Exact matching vs. *WordNet/ConceptNet* matching
- How many concepts are chosen to represent an event?
- To further improve the performance:
  - TF-IDF
  - Term specificity
**Event Search**

- Ranking according to the SQ and concept responses

<table>
<thead>
<tr>
<th><strong>Semantic Query ( q )</strong></th>
<th><strong>Concept Response ( c_i )</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>&lt; Objects &gt;</strong></td>
<td></td>
</tr>
<tr>
<td>Bike</td>
<td>0.60</td>
</tr>
<tr>
<td>Motorcycle</td>
<td>0.60</td>
</tr>
<tr>
<td>Mountain bike</td>
<td>0.60</td>
</tr>
<tr>
<td><strong>&lt; Actions &gt;</strong></td>
<td></td>
</tr>
<tr>
<td>Bike trick</td>
<td>1.00</td>
</tr>
<tr>
<td>Ridding bike</td>
<td>0.62</td>
</tr>
<tr>
<td>Flipping bike</td>
<td>0.61</td>
</tr>
<tr>
<td><strong>&lt; Scenes &gt;</strong></td>
<td></td>
</tr>
<tr>
<td>Parking lot</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Event Search \( s_i = qc_i \)

Video Ranking

* 8000h video
Findings
Findings 1

1. Compared to WordNet/ConceptNet, the simple *exact matching* does the best.

2. The performance is even better by only retaining the *top few* exactly matched concepts.
Findings 1

- Exact matching but only retains the top few concepts
- 7%
Findings 1

Hit the best MAP by only retaining the Top 8 concepts
• Why would only the top few work?

Event 31: Beekeeping

- Bee house (ImageNet)
- Cutting (research collection)
- Cutting down tree (research collection)
- Bee (ImageNet)
- Honeycomb (ImageNet)
• Why would only the top few work?

Event 23: Dog show

Dog show (research collection)

Brush dog (research collection)
• Why ontology-based mapping would not work?

A sample query in TRECVID 2009

```
“find shots of aircraft in sky”
```

visual examples

A sample query in TRECVID 2009

```
“find shots of aircraft in sky”
```

visual examples
• Why ontology-based mapping would not work?

A sample query in TRECVID 2009

- "find shots of aircraft in sky"

Semantic mapping

Visual mapping

Text words

Visual examples
Why ontology-based mapping would not work?

A sample query in TRECVID 2009

```
“find shots of aircraft in sky”
```

Semantic mapping

Visual mapping

Text words

Visual examples
Why ontology-based mapping would not work?

A sample query in TRECVID 2009
Why ConceptNet mapping would not work?

- car
- food
- helmet
- parking lot
- team uniform
- portable shelter

Tailgating
Why *ConceptNet* mapping would not work?

- Tailgating
- Car
- Helmet
- Parking lot
- Team uniform
- Portable shelter
Why *ConceptNet* mapping would not work?

- car
- driver
- engine
- bus
- parking lot
- helmet
- team uniform
- portable shelter
- tailgating
- desires
- at location
- conceptually related to
- has a
• Why ontology-based mapping would not work?

[Diagram showing the concept of "dog" and its mapping issues in ImageNet and SIN.]
Thus, it is difficult to harness the ontology-based mapping while constraining the mapping by event context.

Currently, we only find it useful in:
- Synonyms
  - *E.g.* baby $\rightarrow$ infant
- Strict sub-categories
  - *E.g.* dog $\rightarrow$ husky (哈士奇), german shepherd (德国牧羊犬), ...
  - *hot dog* $\times$
Findings 2

- Lacking concepts?

Human-annotated Concept Sources

- *ImageNet ILSVRC (1000 + 200)*
- *SUN (397)*
- *SIN (346)*
- *Caltech256 (256)*
- *PASCAL VOC (20)*
- *SIN (346)*
- *UCF101 (101)*
- *HMDB51 (51)*
- *HOLLYWOOD2 (22)*
- *Columbia Consumer Video (20)*
- *Olympic Sports (16)*

Added up, the # is still less than 3K
Key concepts may still miss
In the Ad-Hoc event “Extinguishing a Fire”

- Key concepts are missing:
  - Fire extinguisher
  - Firefighter
Findings 2

• Thus, it is reasonable to
  – Scale up the number of concepts, thus increasing the chance of exact match
(1) Outsource concepts

- WikiHow Event Ontology

Yin Cui, Dong Liu, Jiawei Chen, Shih-Fu Chang. **Building A Large Concept Bank for Representing Events in Video.** In *arXiv.*
(2) Learn an embedding space

Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc’Aurelio Ranzato, Tomas Mikolov. 
**DeViSE: A Deep Visual-Semantic Embedding Model.** In *NIPS’13*.

Amirhossein Habibian, Thomas Mensink, Cees G. M. Snoek. 
**VideoStory: A New Multimedia Embedding for Few-Example Recognition and Translation of Events.** In *MM’14*, best paper.
Findings 3

- Improvements by **TF-IDF** and **word specificity**

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP (on MED14-Test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact Matching Only</td>
<td>0.0306</td>
</tr>
<tr>
<td>Exact Matching + TF</td>
<td>0.0420</td>
</tr>
<tr>
<td>Exact Matching + TFIDF</td>
<td>0.0495</td>
</tr>
<tr>
<td>Exact Matching + TFIDF + Word Specificity</td>
<td>0.0502</td>
</tr>
</tbody>
</table>

![Graph with bars showing MAP values for different methods](image)
Outline

- Multimedia Event Detection (MED)
  - Background
  - System Overview
  - Findings

- Multimedia Event Recounting (MER)
  - Background
  - System Workflow
  - Results
Event Recounting

• Summarize a video by evidence localization
  – Given an event query and a test video clip that contains an instance of the event, the system must generate a recounting of the event summarizing the key evidence for the event in the clip. The recounting states:

  – **When**: Intervals of time (or frames) when the event occurred in the clip
  – **Where**: Spatial location in the clip (pixel coordinate or bounding polygon)
  – **What**: A clear, concise textual recounting of the observations
MER System

- In algorithm design, we aim to optimize
  - *Concept-to-event relevancy*
  - *Evidence diversity*
  - *Viewing time of evidential shots*
In algorithm design, we aim to optimize

- **Concept-to-event relevancy**
  - First, we require that candidate shots are *relevant to the event*;
  - Second, we do *concept-to-shot alignment*.

- **Evidence diversity**

- **Viewing time of evidential shots**
In algorithm design, we aim to optimize

- **Concept-to-event relevancy**
  - First, we require that candidate shots are relevant to the event;
  - Second, we do concept-to-shot alignment.

- **Evidence diversity**
  - In concept-to-shot alignment, we recount each shot with a unique concept different from other shots.

- Viewing time of evidential shots
In algorithm design, we aim to optimize

- **Concept-to-event relevancy**
  - First, we require that candidate shots are relevant to the event;
  - Second, we do concept-to-shot alignment.

- **Evidence diversity**
  - In concept-to-shot alignment, we recount each shot with a unique concept different from other shots.

- **Viewing time of evidential shots**
  - Select only the *three* most confident shots as key evidence
  - Basically, each shot is in about *5 seconds*
System Workflow
Key Evidence Localization

Concept Responses

Apply concept detectors
Key Evidence Localization

Choose keyframes/snippets that are most relevant to this event

- All concepts in semantic query are taken into account by calculating the weighted sum $s_i = w r_i$
Key Evidence Localization

The top 3 shots are selected as key evidences
The rests are non-key evidences
• **Concept-to-Shot Alignment**

  The top concept in the key evidence is selected as the representative concept.

  * We choose unique concept for each shot.

  **Semantic Query**

  < Objects >
  - Bike
  - Motorcycle
  - Mountain bike

  < Actions >
  - Bike trick
  - Ridding bike
  - Flipping bike

  < Scenes >
  - Parking lot

The top concept in the key evidence is selected as the representative concept.
Concept-to-Shot Alignment

- The top concept in the key evidence is selected as the representative concept.

* We choose unique concept for each shot.

**Semantic Query**

* **Objects**
  - Bike
  - Motorcycle
  - Mountain bike

* **Actions**
  - Bike trick
  - Ridding bike
  - Flipping bike

* **Scenes**
  - Parking lot

The top concept in the key evidence is selected as the representative concept.*  
* We choose unique concept for each shot.
Concept-to-Shot Alignment

The top concept in the key evidence is selected as the representative concept.

* We choose unique concept for each shot.
**Concept-to-Shot Alignment**

The top concept in the key evidence is selected as the representative concept.

* We choose unique concept for each shot.
• Concept-to-Shot Alignment

The top concept in the key evidence is selected as the representative concept

* We choose unique concept for each shot
Results
Evaluation

View and judge all of the pieces of evidence, then judge how convincing the evidence was as a whole.

Recounting

- **0.99** Attempting a bike trick
  \[(D' \times 0.66) + (S' \times 0.34)\]
- **0.97** Semantic attempting a bike trick

**SUM**

- **0.36** Objects
  \[WEIGHTED\_SUM\]
  - **0.94** bike
  - **0.94** motorcycle
  - **0.30** mountain bike, all terrain bike, off roader

- **0.61** Actions
  \[WEIGHTED\_SUM\]
  - **0.99** bike trick
    - **0.99** bike trick: Visual
  - **0.96** riding bike
    - **0.96** riding bike: Visual
  - **0.96** flipping bike
  - **0.00** skateboard trick

- **0.00** Scenes
  \[WEIGHTED\_SUM\]
  - **0.41** parking lot

---

Evidence

*Press p to play, and r to reset playback.*

[jk] correctly captures the contents of the snippet.

The system chose the right window of time to present the evidence.

The system chose the right bounding box(es) to isolate the evidence.

- **Strongly Agree**
- **Agree**
- **Neutral**
- **Disagree**
- **Strongly Disagree**
Evaluation

View and judge all of the pieces of evidence, then judge how convincing the evidence was as a whole.

Recounting

- 0.99 Attempting a bike trick
  \[(D^*0.66) + (S^*0.34)\]
- 0.97 Semantic attempting a bike trick

**SUM**

- 0.36 Objects
  - 0.94 bike
  - 0.94 bike: Visual
  - 0.94 motorcycle
  - 0.30 mountain bike, all terrain bike, off roader

- 0.61 Actions

**SUM**

- 0.99 bike trick
  - 0.99 bike trick: Visual
  - 0.96 riding bike
  - 0.96 riding bike: Visual

- 0.00 Scenes

**SUM**

- 0.41 parking lot

Evidence

Press p to play, and r to reset playback.

[riding bike] correctly captures the contents of the snippet.

The system chose the right window of time to present the evidence.

The system chose the right bounding box(es) to isolate the evidence.

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

Submit
View and judge all of the pieces of evidence, then judge how convincing the evidence was as a whole.

**Recounting**

- 0.98 Attempting a bike trick
  \( (D*0.66) + (S*0.34) \)
- 0.96 Semantic attempting a bike trick
  \( SUM \)
  - 0.37 Objects
    \( WEIGHTED\_SUM \)
    - 0.99 bike
      - 0.99 bike: Visual
    - 0.99 motorcycle
      - 0.99 motorcycle: Visual
    - 0.48 mountain bike, all terrain bike, off roader
- 0.59 Actions
  \( WEIGHTED\_SUM \)
  - 0.99 bike trick
    - 0.99 bike trick: Visual
  - 0.97 riding bike
  - 0.84 flipping bike
  - 0.61 skateboard trick
- 0.00 Scenes
  \( WEIGHTED\_SUM \)
  - 0.30 parking lot

**Evidence**

Press p to play, and r to reset playback.

- [motorcycle] correctly captures the contents of the snippet.
- The system chose the right window of time to present the evidence.
- The system chose the right bounding box(es) to isolate the evidence.

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

**Submit**
MER14 Results

The percentage of strongly agree

(a) Evidence quality
(b) Event query quality
The percentage of both agree and strongly agree
Summary
Summary

- Zero-Example MED System
  - A good baseline: the simple *exact matching* shows reasonable performance
  - Don’t include noisy concepts
  - The *context* of a concept is important in event detection. Only referring to the name is insufficient. It remains a problem of how to well combine the event context into concept knowledge base.
MER System

- In key evidence localization, we emphasize the event relevancy first, then the hot concepts.
- We recommend three shots as key evidences and each in about 5 seconds.
Thanks for your attention!