Concept-Based Framework for Detecting High-Level Events in Video

Shih-Fu Chang
Columbia University

September 18th, 2014
University of Maryland

Acknowledgments: This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20071. 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 express or implied, of IARPA, DoI/NBC, or the U.S. Government.
Complex Video Event Detection

What “event” is described in these videos?

Felling a Tree  
Attempting Board Trick  
Horseriding Competition

Automatic Detection will help:

High-Level Summarization

Targeted Advertisement/Learning
Recognize Complex Events

- NIST TRECVID Multimedia Event Detection (MED) contest
- Detecting complex events in ~200,000 videos
High-level Events are Complex

- Y. Jiang et al., high level events recognition in unconstrained videos, IJMIR, 2012 (survey)
Decompose an event into concepts

Event
- Action
  - Lighting Candles
  - Cutting Cake
  - Drinking
  - Dancing
- Scene
  - Home
  - Restaurant

Birthday Party
- Action
  - Cake
  - Candle
  - Candy
  - Gift
  - Food
- Scene
  - Park
Concepts as Mid-Level Representation for Video

Playing Bike Trick

SIFT LBP Trajectories Gist

Audio

Mid-level concept pool

Zero-Shot Learning (without training samples)

Statistical, spatio-temporal

Tree Dog Bike Sky Bike Trick Run Grass Pizza Person Ride Road Cat Wheelie...
How to Discover Relevant Concepts?

A question often skipped by computer vision community.

Some work in attribute discovery, but not for video events.

Ways of discovering new complex event concepts:

- Use dictionary definitions
- Discover concepts from the Web
- Explore knowledge source, e.g., *WikiHow* events
- Perform interactive cognitive studies
Dictionary: TRECVID MED Event Kits

Groom An Animal

• **Definition:** One or more people groom an animal.

• **Explication:** Grooming refers to caring for the hygiene/cleanliness and appearance of the animal. A very common form of grooming is *bathing* the animal, usually accomplished by either immersing the animal in water or spraying the animal with water, often followed by *application* of soap/shampoo and then additional rinsing with water. Other grooming activities include *trimming* of hair and nails, cleaning of teeth, eyes, and ears, and *brushing, combing, and styling* the fur of the animal.

• **Evidence Description:**
  - **Scene:** yard, corral, bathroom, grooming salon, exhibition center.
  - **Object/people:** sink, bathtub, hose, shower, soap, shampoo, scissors, clippers.
  - **Activity:** spraying hose, putting animal on table, rinsing, blow drying fur, cutting fur, clipping nails.

Obviously hard to scale up!
How to Discover Relevant Concepts?

- Use dictionary definitions
- **Discover Concepts from the Web**
- Use knowledge source like WikiHow
- Perform interactive cognitive studies
Discover Concepts via Web Data Expansion

Supervised Event Modeling
Zero-Shot Event Retrieval
Video Semantic Recounting

Query

flickr

Noisy tag filtering

Concept Classifier

dog

groom

haircut

horse

2012

sunny

girl

cat: 0.01, horse: 0.01, person: 0.8, grooming: 0.7, dog: 0.9, grass: 0.9, ...

cat: 0.9, horse: 0.01, person: 0.9, grooming: 0.8, dog: 0.02, grass: 0.01, ...

Concept based video representation

Jiawei Chen, Yin Cui, Guangnan Ye, Dong Liu, Shih-Fu Chang. Event-Driven Semantic Concept Discovery by Exploiting Weakly Tagged Internet Images. In ACM International Conference on Multimedia Retrieval (ICMR), full paper (oral), 2014.
Concepts discovered from Web

(Attempting a bike trick)
# Web-Concept Expansion Finds Novel Concepts

<table>
<thead>
<tr>
<th>Event Name</th>
<th>Concepts Discovered from Different Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>getting a vehicle unstuck</td>
<td><code>Classemes</code> air transportation vehicle, all terrain vehicle, amphibious vehicle, armed person, armored fighting vehicle, armored recovery vehicle, armored vehicle, armored vehicle heavy, armored vehicle light, command vehicle</td>
</tr>
<tr>
<td></td>
<td><code>ImageNet</code> [vehicle, bumper car, craft, military vehicle, rocket, skibob, sled, steamroller, wheeled vehicle, conveyance]</td>
</tr>
<tr>
<td></td>
<td><code>Web</code> tire, car, snow, stick, stuck, winter, vehicle, truck, night, blizzard</td>
</tr>
<tr>
<td>grooming an animal</td>
<td><code>Classemes</code> adult animal, animal, animal activity, animal blo, animal body part, animal body region, animal cage, animal container, animal pen, animal shelter</td>
</tr>
<tr>
<td></td>
<td><code>ImageNet</code> groom, animal, invertebrate, homeotherm, work animal, darter, range animal, creepy-crawly, domestic animal, molter</td>
</tr>
<tr>
<td></td>
<td><code>Web</code> dog, pet, grooming, cat, animal, bath, cute, canine, puppy, water</td>
</tr>
<tr>
<td>making a sandwich</td>
<td><code>Classemes</code> baking dish, cafe place, classroom setting, collection display setting, cutting device, dish drying rack, food utensil, hair cutting razor, hdtv set, hole making tool</td>
</tr>
<tr>
<td></td>
<td><code>ImageNet</code> sandwich, open-face sandwich, butty, reuben, ham sandwich, gyro, chicken sandwich, hotdog, club sandwich, wrap</td>
</tr>
<tr>
<td></td>
<td><code>Web</code> sandwich, food, bread, cooking, cheese, spice, baking, pan, kitchen, breakfast</td>
</tr>
</tbody>
</table>

Top 10 concepts discovered from different resources
Concept images are noisy

- Training Image Selection

Images annotated with “dog”

Kernel Density Estimation (KDE)

Sampling from the dense region in the feature space.

Rank all images based on their confidence scores derived from KDE.

clean dog images

Noisy images or outliers

SVM with RBF kernel

Concept classifier for dog
Event-specific expansion finds relevant images

<table>
<thead>
<tr>
<th></th>
<th>“vehicle” in event “getting a vehicle unstuck”</th>
<th>“dog” in event “dog show”</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CC</strong></td>
<td><img src="image1" alt="Vehicle images" /></td>
<td><img src="image2" alt="Dog images" /></td>
</tr>
<tr>
<td><strong>IN</strong></td>
<td><img src="image3" alt="Vehicle images" /></td>
<td><img src="image4" alt="Dog images" /></td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><img src="image5" alt="Vehicle images" /></td>
<td><img src="image6" alt="Dog images" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>“groom” in “grooming an animal”</th>
<th>“ring” in “marriage proposal”</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CC</strong></td>
<td><img src="image7" alt="Groom images" /></td>
<td><img src="image8" alt="Ring images" /></td>
</tr>
<tr>
<td><strong>IN</strong></td>
<td><img src="image9" alt="Groom images" /></td>
<td><img src="image10" alt="Ring images" /></td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><img src="image11" alt="Groom images" /></td>
<td><img src="image12" alt="Ring images" /></td>
</tr>
</tbody>
</table>

CC: Classemes IN: ImageNet
Human Evaluation

• For each event, choose top 100 concepts based on similarity between concept name and event name.

• Our concept is better than others with **81.29%** chance.
• ImageNet concept is better than others with **34.19%** chance.
• Classemes concept is better than others with **32.46%** chance.

Choose one event

Choose 5 concepts @ same ranks

Choose 5 concepts @ same ranks

Pair-wise comparison

Which 5 concepts are more relevant to event?

Repeat 1000 times
Zero-Shot Event Retrieval

- Given an event name as textual query without any training videos, rank all videos.
- A simple search method using concepts

\[
\text{Ranking Score} = \sum_{i=1}^{T} \text{sim}(\text{concept}_i, \text{Event}) \times \text{score}_i
\]

\(T: \# \text{ of concepts}\)
Experiment Setup

• 20 Pre-specified MED events
  – No training video for each event.
  – Use 100 dimension concept score vector for each event as feature representation.
  – Evaluation metric: full-length Mean Average Precision (MAP).
Performance: Zero-Shot Concept-based Event Retrieval

CC: Classeme concepts, CIN: ImageNet, CRI/CSI: Ours

How to Discover Relevant Concepts?

• Use dictionary definitions
• Discover Concepts from the Web
• **Use knowledge source like WikiHow**
• Perform interactive cognitive studies
Knowledge Source: WikiHow

- A wiki contains ~180,000 articles on 2,803 “how to” categories.
- All articles are organized into a hierarchical structure.
Construct **EventNet from WikiHow**

Article Titles
- How to train a horse to recognize commands
- How to make a basket from a garden hose
- Extract “direct verb-object pair” from article titles
- Train horse, make basket, ……

WikiHow Hierarchical Ontology

All direct verb-object pairs
- calculate frequency of each pair
- Assess visual detectability of each pair

699 visually detectable events

Discover visual concepts from articles or Web

Events and their specific concepts
- groom animals: dog, cat, water, groom, brush, scissor, ….
- play soccer: soccer, goal, filed, player, run, ….
- bee keeping: bee, flower, case, farm, people, honey ….

Total about 12,000 concepts

Y. Cui, et al, Building A Large Concept Bank for Representing Events in Video, arXiv, March 2014
Columbia EventNet Browser
How to Discover Relevant Concepts?

- Use dictionary definitions
- Discover Concepts from the Web
- Use knowledge source like *WikiHow*
- Perform interactive cognitive studies
How Do Humans Judge Events?

Conventional: Passive Linear Video Playback and Judge

- Used extensively in video summarization [1,2], persistent surveillance [3].

Human Decision Strategy

• Do humans actually employ linear playback to judge a complex event?
  • Probably not!

• Bubble-game [5] to identify “Human” discovered discriminative patches for fine-grained recognition

Evidences needed by humans for event judgment?

Look for **Needed Evidence** in Events **(Proposed)**

- Not necessarily linear
- Not necessarily sequential

Let’s take a test ...

Do the videos depict Cleaning an appliance? Lose 2 point for each additional hint.
Lets take a test ...

Do the videos depict **Cleaning an appliance**? Lose 2 point for each additional hint.
Let's take a test ...

Do the videos depict **Cleaning an appliance**? Lose 2 point for each **additional hint**.
**Let's take a test ...**

Do the videos depict **Cleaning an appliance**? Lose 2 point for each additional hint.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
</tbody>
</table>

8 points

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
</tbody>
</table>

12 points
**Let's take a test ...**

Do the videos depict **Cleaning an appliance**? Lose 2 point for each **additional hint**.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Lets take a test ...**

Do the videos depict **Cleaning an appliance**? Lose 2 point for each **additional hint**.

<table>
<thead>
<tr>
<th></th>
<th>Reveal?</th>
<th>Reveal?</th>
<th>Reveal?</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Green Checkmark]</td>
<td>![Green Checkmark]</td>
<td>![Green Checkmark]</td>
<td>![Green Checkmark]</td>
</tr>
<tr>
<td>![Red X]</td>
<td>![Red X]</td>
<td>![Red X]</td>
<td>![Red X]</td>
</tr>
</tbody>
</table>

8

8
**Let's take a test ...**

Do the videos depict **Cleaning an appliance**? Lose 2 point for each **additional hint**.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image1.png" alt="Image 1" /></td>
<td><img src="image2.png" alt="Image 2" /></td>
<td><img src="image3.png" alt="Image 3" /></td>
<td><img src="image4.png" alt="Image 4" /></td>
<td><img src="image5.png" alt="Image 5" /></td>
<td><img src="image6.png" alt="Image 6" /></td>
<td><img src="image7.png" alt="Image 7" /></td>
<td><img src="image8.png" alt="Image 8" /></td>
<td><img src="image9.png" alt="Image 9" /></td>
<td><img src="image10.png" alt="Image 10" /></td>
<td><img src="image11.png" alt="Image 11" /></td>
<td><img src="image12.png" alt="Image 12" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td><img src="image13.png" alt="Image 13" /></td>
<td><img src="image14.png" alt="Image 14" /></td>
<td><img src="image15.png" alt="Image 15" /></td>
<td><img src="image16.png" alt="Image 16" /></td>
<td><img src="image17.png" alt="Image 17" /></td>
<td><img src="image18.png" alt="Image 18" /></td>
<td><img src="image19.png" alt="Image 19" /></td>
<td><img src="image20.png" alt="Image 20" /></td>
<td><img src="image21.png" alt="Image 21" /></td>
<td><img src="image22.png" alt="Image 22" /></td>
<td><img src="image23.png" alt="Image 23" /></td>
<td><img src="image24.png" alt="Image 24" /></td>
</tr>
<tr>
<td></td>
<td><img src="image25.png" alt="Image 25" /></td>
<td><img src="image26.png" alt="Image 26" /></td>
<td><img src="image27.png" alt="Image 27" /></td>
<td><img src="image28.png" alt="Image 28" /></td>
<td><img src="image29.png" alt="Image 29" /></td>
<td><img src="image30.png" alt="Image 30" /></td>
<td><img src="image31.png" alt="Image 31" /></td>
<td><img src="image32.png" alt="Image 32" /></td>
<td><img src="image33.png" alt="Image 33" /></td>
<td><img src="image34.png" alt="Image 34" /></td>
<td><img src="image35.png" alt="Image 35" /></td>
<td><img src="image36.png" alt="Image 36" /></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image37.png" alt="Image 37" /></td>
<td><img src="image38.png" alt="Image 38" /></td>
<td><img src="image39.png" alt="Image 39" /></td>
<td><img src="image40.png" alt="Image 40" /></td>
<td><img src="image41.png" alt="Image 41" /></td>
<td><img src="image42.png" alt="Image 42" /></td>
<td><img src="image43.png" alt="Image 43" /></td>
<td><img src="image44.png" alt="Image 44" /></td>
<td><img src="image45.png" alt="Image 45" /></td>
<td><img src="image46.png" alt="Image 46" /></td>
<td><img src="image47.png" alt="Image 47" /></td>
<td><img src="image48.png" alt="Image 48" /></td>
<td><img src="image49.png" alt="Image 49" /></td>
</tr>
<tr>
<td></td>
<td><img src="image50.png" alt="Image 50" /></td>
<td><img src="image51.png" alt="Image 51" /></td>
<td><img src="image52.png" alt="Image 52" /></td>
<td><img src="image53.png" alt="Image 53" /></td>
<td><img src="image54.png" alt="Image 54" /></td>
<td><img src="image55.png" alt="Image 55" /></td>
<td><img src="image56.png" alt="Image 56" /></td>
<td><img src="image57.png" alt="Image 57" /></td>
<td><img src="image58.png" alt="Image 58" /></td>
<td><img src="image59.png" alt="Image 59" /></td>
<td><img src="image60.png" alt="Image 60" /></td>
<td><img src="image61.png" alt="Image 61" /></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Lets take a test ...**

Do the videos depict *Cleaning an appliance*? Lose 2 point for each *additional hint*.

|---|---------|---------|---------|---------|---------|---------|---------|---------|
**Lets take a test ...**

Do the videos depict *Cleaning an appliance*? Lose 2 point for each *additional hint*.

<table>
<thead>
<tr>
<th>8</th>
<th>Reveal?</th>
<th>Reveal?</th>
<th>Reveal?</th>
</tr>
</thead>
<tbody>
<tr>
<td>✅</td>
<td><img src="image1.png" alt="Image 1" /></td>
<td><img src="image2.png" alt="Image 2" /></td>
<td><img src="image3.png" alt="Image 3" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8</th>
<th>Reveal?</th>
<th>Reveal?</th>
<th>Reveal?</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td><img src="image4.png" alt="Image 4" /></td>
<td><img src="image5.png" alt="Image 5" /></td>
<td><img src="image6.png" alt="Image 6" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>10</th>
<th>Reveal?</th>
<th>Reveal?</th>
<th>Reveal?</th>
<th>Reveal?</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td><img src="image7.png" alt="Image 7" /></td>
<td><img src="image8.png" alt="Image 8" /></td>
<td><img src="image9.png" alt="Image 9" /></td>
<td><img src="image10.png" alt="Image 10" /></td>
</tr>
</tbody>
</table>
Let's take a test ...

Do the videos depict Cleaning an appliance? Lose 2 point for each additional hint.

|---------|---------|---------|---------|---|

Congratulations you got all correct! You scored 26 out of 30.
# Minimally Needed Evidence

For the event **Cleaning an appliance**:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>![Green Check]</td>
<td>![Number 1]</td>
<td>![Number 2]</td>
</tr>
</tbody>
</table>

- **Practical way** of finding Minimally Needed Evidence
  → **Event Quiz Interface**

- **Clever annotation tool** → Enables *judicious use* of Human feedback

- Can **reduce computational overhead** for feature extraction
Microshot Selection

• Most action concepts e.g. jogging, boxing, can be captured in 1.5s of continuous footage (30hz)

• Divide video into non-overlapping 1.5s blocks; Filter out non-interesting microshots (low appearance + motion entropy)
Surprisingly, Humans can

- Correctly judge an event in ~87% cases from just 1 or 2 microshots (1.5s footage)
Surprisingly, Humans can

- **Correctly identify** a video containing an event in videos in 55% cases with 1 shot
- **Correctly reject** a video for not containing an event in 68% cases
Additionally- Event Complexity varies

- “Tuning a musical instrument” more visually challenging than “Fixing musical instrument” (needs more microshot revelation than other events)
Additionally- Event Complexity varies

- “Tailgating”, “Horseriding Competition” and “Beekeeping” require less evidence
Hypothesis

• Human Discovered Evidences provide better event representation for recognition

• To validate: Basic Retrieval experiment
Representation for Retrieval

- Standard Bag-of-visual Words approach
- Empirically determined vocabulary size for
  - Appearance Features: 2,000
Representation for Retrieval

- Standard Bag-of-visual Words approach
- Empirically determined vocabulary size for
  - Motion Features: 5,000
Retrieval Methodology

Target Event: Cleaning an Appliance

Top videos for Cleaning an Appliance (Likely)
Use Negative Cues for Quick Reject

Target Event: **Cleaning an Appliance**

Top-3 Negative Clusters for “Cleaning an Appliance”

Train 1-class SVM

1-class model to filter non (**Cleaning an Appliance**)
Retrieval Methodology

Target Event: Cleaning an Appliance

Top videos for Cleaning an Appliance (Likely)

Top videos for Changing an Appliance (Likely)

1-class model to filter non (Cleaning an Appliance)

Refined Ranked List of Results

- High Similarity
- Moderate Similarity
- Low Similarity
- Ground Truth Positive
- Ground Truth Negative
- Order of Revelation
Rewarding Decisive Microshots

Target Event: Cleaning an Appliance

Reveal?

Reveal?

Reveal?

Positives for Cleaning an Appliance (Likely)
Rewarding Decisive Microshots

Target Event: Cleaning an Appliance

\[ v_i^{(j)} = (N - Q + i) \times \exp(-|x_i^{(j)} - m_i|) \]
Rewarding Decisive Microshots

Target Event: Cleaning an Appliance

\[ u_i^{(j)} = (N - Q + i) \times \exp\left(-|x_i^{(j)} - m_i|\right) \]
## Experiments

<table>
<thead>
<tr>
<th>ID</th>
<th>Event Name</th>
<th>[N]</th>
<th>ID</th>
<th>Event Name</th>
<th>[N]</th>
</tr>
</thead>
<tbody>
<tr>
<td>E006</td>
<td>Birthday</td>
<td>173</td>
<td>E007</td>
<td>Changing Tire</td>
<td>111</td>
</tr>
<tr>
<td>E008</td>
<td>Flash-mob</td>
<td>173</td>
<td>E009</td>
<td>Vehicle Unstuck</td>
<td>132</td>
</tr>
<tr>
<td>E010</td>
<td>Grooming Animal</td>
<td>138</td>
<td>E011</td>
<td>Making Sandwich</td>
<td>126</td>
</tr>
<tr>
<td>E012</td>
<td>Parade</td>
<td>138</td>
<td>E013</td>
<td>Parkour</td>
<td>112</td>
</tr>
<tr>
<td>E014</td>
<td>Repairing Appl.</td>
<td>123</td>
<td>E015</td>
<td>Sewing Project</td>
<td>120</td>
</tr>
<tr>
<td>E021</td>
<td>Bike-trick</td>
<td>200</td>
<td>E022</td>
<td>Giving Directions</td>
<td>200</td>
</tr>
<tr>
<td>E023</td>
<td>Dog-show</td>
<td>200</td>
<td>E024</td>
<td>Wedding</td>
<td>200</td>
</tr>
<tr>
<td>E025</td>
<td>Marriage Proposal</td>
<td>200</td>
<td>E026</td>
<td>Renovating Home</td>
<td>200</td>
</tr>
<tr>
<td>E027</td>
<td>Rock-climbing</td>
<td>200</td>
<td>E028</td>
<td>Town-hall Meet</td>
<td>200</td>
</tr>
<tr>
<td>E029</td>
<td>Winning Race</td>
<td>200</td>
<td>E030</td>
<td>Metal crafts</td>
<td>200</td>
</tr>
</tbody>
</table>

- NIST Multimedia Event Detection TEST Dataset 2011-12 : 20 Events
- NIST Multimedia Event Detection ADHOC Dataset 2013 : 10 Events
**Baselines**

- **Baselines (BL):**
  - A. Use all microshots
  - B. Use automatic microshot selection using scene aligned pooling [7]

### Retrieval Results

<table>
<thead>
<tr>
<th>Events</th>
<th>BL-A</th>
<th>BL-B</th>
<th>MNE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beekeeping</td>
<td>3.47</td>
<td>4.12</td>
<td>20.96</td>
</tr>
<tr>
<td>Wedding shower</td>
<td>2.87</td>
<td>2.05</td>
<td>17.23</td>
</tr>
<tr>
<td>Non-motorized Vehicle repair</td>
<td>2.56</td>
<td>3.35</td>
<td>16.90</td>
</tr>
<tr>
<td>Fixing musical instrument</td>
<td>3.52</td>
<td>3.09</td>
<td>19.26</td>
</tr>
<tr>
<td>Horse riding competition</td>
<td>4.60</td>
<td>5.21</td>
<td>21.46</td>
</tr>
<tr>
<td>Felling a tree</td>
<td>5.47</td>
<td>5.25</td>
<td>20.86</td>
</tr>
<tr>
<td>Parking a vehicle</td>
<td>3.09</td>
<td>6.11</td>
<td>17.04</td>
</tr>
<tr>
<td>Playing fetch</td>
<td>2.73</td>
<td>4.08</td>
<td>16.62</td>
</tr>
<tr>
<td>Tailgating</td>
<td>1.75</td>
<td>3.15</td>
<td>15.48</td>
</tr>
<tr>
<td>Tuning musical instrument</td>
<td>3.95</td>
<td>4.06</td>
<td>18.26</td>
</tr>
<tr>
<td>Mean Average Precision</td>
<td>3.41</td>
<td>4.07</td>
<td>18.47</td>
</tr>
</tbody>
</table>

- MED13 ADHOC data set
- Only using MNE, absolute performance gain ~ 14%
### Retrieval Results

<table>
<thead>
<tr>
<th>Events</th>
<th>BL-A</th>
<th>BL-B</th>
<th>MNE</th>
<th>MNE+QR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beekeeping</td>
<td>3.47</td>
<td>4.12</td>
<td>20.96</td>
<td>20.86</td>
</tr>
<tr>
<td>Wedding shower</td>
<td>2.87</td>
<td>2.05</td>
<td>17.23</td>
<td>17.42</td>
</tr>
<tr>
<td>Non-motorized Vehicle repair</td>
<td>2.56</td>
<td>3.35</td>
<td>16.90</td>
<td>17.09</td>
</tr>
<tr>
<td>Fixing musical instrument</td>
<td>3.52</td>
<td>3.09</td>
<td>19.26</td>
<td>19.69</td>
</tr>
<tr>
<td>Horse riding competition</td>
<td>4.60</td>
<td>5.21</td>
<td>21.46</td>
<td>21.91</td>
</tr>
<tr>
<td>Felling a tree</td>
<td>5.47</td>
<td>5.25</td>
<td>20.86</td>
<td>21.27</td>
</tr>
<tr>
<td>Parking a vehicle</td>
<td>3.09</td>
<td>6.11</td>
<td>17.04</td>
<td>17.35</td>
</tr>
<tr>
<td>Playing fetch</td>
<td>2.73</td>
<td>4.08</td>
<td>16.62</td>
<td>16.74</td>
</tr>
<tr>
<td>Tailgating</td>
<td>1.75</td>
<td>3.15</td>
<td>15.48</td>
<td>14.97</td>
</tr>
<tr>
<td>Tuning musical instrument</td>
<td>3.95</td>
<td>4.06</td>
<td>18.26</td>
<td>18.56</td>
</tr>
<tr>
<td><strong>Mean Average Precision</strong></td>
<td>3.41</td>
<td>4.07</td>
<td>18.47</td>
<td>18.59</td>
</tr>
</tbody>
</table>

- Only using MNE, absolute performance gain ~ 14%
- Quick Rejection does not drastically improve the performance
- But can significantly speedup
Retrieval with Minimal Needed Evidences Achieves 4X Performance

- Cp. baselines: (A) all shots (B) scene clusters (Cao, et al, ECCV12)
What Concepts are Critical for Human Judgement?

• Discover needed concepts for humans
• Concepts those cannot be found from textual description of an event
Birthday Party

Positive Concepts: “Yes, because I see ...”

- balloon
- cake
- candle
- chair
- clapping
- dining_table
- food
- gift
- group_of_people
- indoor
- party_hat
- person
- singing
- smiling_face
- table
- wine

Must-Have Concepts: “No, because I don’t see ...”

- balloon
- cake
- candle
- person

Concepts parsed from Textual Event Kit

- birthday
- balloon
- outdoor
- park
- streamer
- anniversary
- cake
- candle
- restaurant
- celebration
- children
- host
- conical
- blowing
- party
- indoor
- lit
- shiny/colorful
- singing
- food
- game
- gift
- guest
- home
- honor
Repairing an Appliance

Positive Concepts: “Yes, because I see …”

Must-Have Concepts: “No, because I don’t see …”

Concepts parsed from Textual Event Kit
Minimal Evidence Approach

• Discovers unique concepts needed for humans
• Must-Have concepts can be used for quick rejection
• Positive concepts can be used for efficient detection
Does Minimal Evidence help Event Recognition?

Positive Videos

Where are evidences?

Negative Videos

Learn Key Evidences

1. Decompose videos into static & dynamic instances

- Static instances (SIFT BoW frames)
- Dynamic instances (3, 5, 10, 15, 20 secs MBH video clips)

2. Learn static/dynamic instance codebook by k-means

3. Create instance codewords

- Static instances: $Cs = \{c_{s1}, c_{s2}, c_{s3}, \ldots\}$
- 5-sec dynamic instances: $Cd = \{c_{d1}, c_{d2}, c_{d3}, \ldots\}$
- 15-sec dynamic instances: $C_{d15} = \{c_{d15,1}, c_{d15,2}, c_{d15,3}, \ldots\}$
Embed video instances into codewords via max similarity measure:

\[ s(S_i, c^s_l) = \text{max pooling from video } S_i \text{ to codeword } c^s_l \]

\[ s(S_i, c^s_l) = \max_{1 \leq j \leq n_i} \left( \exp \left( - \frac{d(s_{ij}, c^s_l)}{\sigma} \right) \right) \]

embedded representation for video \( S \)

\[ m_i = \left[ s(S_i, c^s_1), \ldots s(S_i, c^s_{G_S}), s(D_i, c^d_1), \ldots s(D_i, c^d_{G_d}) \right]^T \]
\[
\min_w \|w\|_1 + \lambda \max_{1 \leq n \leq N^-} \left( \frac{1}{N^+} \sum_{p=1}^{N^+} \max(1 - w^T (m^+_p - m^-_n), 0) \right)
\]

- Infinite push loss (Agarwal, 2011) used to push positive videos to the top ranks
- Other ranking orders ignored
- \(L_1\) norm used to select key codewords and evidences
- Optimize with Alternating Direction Method of Multipliers
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Birthday party</td>
<td>2.25%</td>
<td>4.38%</td>
<td>6.08%</td>
<td>6.24%</td>
<td>5.08%</td>
<td>7.45%</td>
</tr>
<tr>
<td>Change a vehicle tire</td>
<td>0.76%</td>
<td>0.92%</td>
<td>3.96%</td>
<td>24.62%</td>
<td>9.50%</td>
<td>14.44%</td>
</tr>
<tr>
<td>Flash mob gathering</td>
<td>8.30%</td>
<td>15.29%</td>
<td>35.28%</td>
<td>37.46%</td>
<td>33.77%</td>
<td>40.87%</td>
</tr>
<tr>
<td>Get a vehicle. unstuck</td>
<td>1.95%</td>
<td>2.04%</td>
<td>8.45%</td>
<td>15.72%</td>
<td>7.38%</td>
<td>7.72%</td>
</tr>
<tr>
<td>Groom an animal</td>
<td>0.74%</td>
<td>0.74%</td>
<td>3.05%</td>
<td>2.09%</td>
<td>1.76%</td>
<td>1.83%</td>
</tr>
<tr>
<td>Make a sandwich</td>
<td>1.48%</td>
<td>0.84%</td>
<td>4.95%</td>
<td>7.65%</td>
<td>3.13%</td>
<td>4.86%</td>
</tr>
<tr>
<td>Parade</td>
<td>2.65%</td>
<td>4.03%</td>
<td>8.95%</td>
<td>12.01%</td>
<td>14.34%</td>
<td>17.69%</td>
</tr>
<tr>
<td>Parkour</td>
<td>2.05%</td>
<td>3.04%</td>
<td>24.62%</td>
<td>10.96%</td>
<td>20.14%</td>
<td>25.3%</td>
</tr>
<tr>
<td>Repair an appliance</td>
<td>4.39%</td>
<td>10.88%</td>
<td>19.81%</td>
<td>32.67%</td>
<td>25.81%</td>
<td>31.75%</td>
</tr>
<tr>
<td>Work on sewing project</td>
<td>0.61%</td>
<td>5.48%</td>
<td>6.53%</td>
<td>7.49%</td>
<td>4.66%</td>
<td>8.34%</td>
</tr>
<tr>
<td>mean AP</td>
<td>2.52%</td>
<td>4.77%</td>
<td>12.27%</td>
<td>15.69%</td>
<td>12.56%</td>
<td>16.02%</td>
</tr>
</tbody>
</table>
Another Formulation: Multiple Instance Learning

Positive Videos

Negative Videos
MIL assumptions not right

• Negative videos may still have partial evidences
• Positive videos should have more (critical) evidence proportions
Relax MIL Setting – Soft Bag Proportions

Positive Videos

75%

Learning optimal proportions of bags

Negative Videos

14%

83%

Videos
The $\alpha$SVM Algorithm

F. Yu; D. Liu; S. Kumar; T. Jebara; S.-F. Chang. $\alpha$SVM for learning with label proportions. ICML13

- Large-margin framework:

$$\min_{\mathbf{w}, b} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{N} L(y_i, \mathbf{w}^T \varphi(x_i) + b) + C' \sum_{k=1}^{K} L_p(\tilde{p}_k(y), p_k)$$

s.t. $\forall_{i=1}^{N}, y_i \in \{-1, 1\}$.

- Generalizes the classic SVM.
- Naturally spans supervised/semi-supervised learning and clustering.
- Learned with alternate optimization or a relaxed convex form

Microshot as instances
Video as Bags, Learn label proportion for each bag
Dealing with Unknown Proportion

• But bag proportion is unknown a priori
• Set initial proportion \( P_m \) to 1 (0) for positive (negative) videos

\[
\min_{\{y^m\}_{m=1}^M, w, b} \frac{1}{2}||w||^2 + C \sum_{m=1}^M \sum_{i=1}^{N_m} L(y_i^m, (w^\top x_i^m + b)) \\
+ C_p \sum_{m=1}^M |p_m(y^m) - P_m|
\]

\[
s.t. \quad P_m = \begin{cases} 
1 & \text{if } Y_m = 1 \\
0 & \text{if } Y_m = -1 
\end{cases} , m = 1, \ldots, M.
\]

Infer latent variables: bag proportion and instance labels
Learn SVM classifier \( w \)
Optimization Procedure

• Alternating optimization for p-SVM
  1. Fix instance labels $y$ and learn $w$ and $b$. The problem becomes a classic SVM

  2. Fix $w$ and $b$ and update instance labels $y$, calculate positive instance proportions $p(y)$

  3. Optimize proportions of videos independently
    3-1. Compute all possible loss values (prediction loss + proportion loss) by flipping instance labels one by one
    3-2. Sort all loss values to find optimal instance proportion
Complexity Analysis

• Running time of one iteration
  = SVM training time + instance sorting time
  = $O(N + I_{max} \log I_{max})$, $I_{max}$ = max number of instances in a video

• SVM training time $>>$ instance sorting time, so the complexity is the same as SVM
## Also Learn Best Granularities for Event Evidences

<table>
<thead>
<tr>
<th>Optimal Granularity</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 sec (MBH)</td>
<td>none</td>
</tr>
<tr>
<td>15 sec (MBH)</td>
<td>1. Attempting board trick; 15. Sewing project</td>
</tr>
<tr>
<td>20 sec (MBH)</td>
<td>20. Marriage proposal</td>
</tr>
</tbody>
</table>
Experimental Results

- 20% performance gain on MED11 dataset

- 10% performance gain on MED12 dataset
Rock Climbing

Detected Evidences

Optimal granularity: 1 frame
Rock Climbing

Detected Evidences

Optimal granularity: 1 frame
Winning Race without a Vehicle

Detected Evidences

Optimal granularity: 1 frame
Winning Race without a Vehicle
Detected Evidences for Dog Show

Optimal granularity:
3 seconds
Conclusions

- **Concepts as Event mid-level representation**
- **Questions:**
  - How to find/combine them?
  - How to train classifiers?
  - How to locate evidences in videos?
  - How to incorporate temporal dynamics?
- **Tools:**
  - Preparing EventNet
  - About 1,000 events & ~12,000 Concepts


