RECOGNIZING COMPLEX EVENTS IN VIDEOS

Amirhossein Habibian and Cees G. M. Snoek
University of Amsterdam
Goal

Interaction of people and objects at a specific place
Recognizing any event that happens in a video

Rock climbing  Wedding ceremony  Cleaning an appliance
Event Recognition

We focus on representing video events

Videos → Event Representation → Event Modeling

Positive examples
Negative examples

Supervised Learning
Related Work:

Low-level Representation

Events as histogram of low-level features

- Image features: SIFT variations, deep learning, etc.
- Audio features: MFCC, AUD, etc.
- Text features: ASR, OCR, etc.
- Motion features: STIP, dense trajectories, etc.

Good recognition accuracy, no interpretation

[P. Natarjan et al., CVPR12] [Tamarkar et al., CVPR12] and many others
Related Work:

Semantic Representation

Events as histogram of concept detector responses

[Merler, TMM12][Althoff, ACM-MM12][Li, NIPS10][Mazloom, ICMR13]
Contribution

- We study concept vocabulary for representing events

Research Questions

1. How many concepts to include?
2. What concept types to include?
3. How accurate should the concept detectors be?
Inspiration: Human Event Description

A woman folds and packages a scarf she has made.

People competing in a sand sculpting competition and children playing on the beach.

- We analyze 13K web videos and their descriptions
Human Concept-Vocabulary

- Consists of 5K distinct and mostly rare concepts
- It is composed of various concept types

Bar chart showing portions of categories including Animal, People, Action, Scene, Attribute, and Non Visual.
Experimental Setup

- **Event Dataset**
  - Based on TRECVID Multimedia Event Detection 2012
  - 13,274 Videos (66% Training)
  - 25 Events i.e. renovating home and making sandwich
  - Average Precision

- **Vocabulary sampled from 1,346 concept detectors**
  - Data from ImageNet Challenge’11 and TRECVID SIN’12
  - Fisher coding of SIFT, OpponentSIFT and RGB-SIFT
  - Linear SVM
All concepts categorized by type

- **Object**
  - Helicopter
  - Tank
  - Bus
  - Canoe
  - Harmonica
  - Boat ship
  - Bicycle
  - Chair
  - Cell phone
  - Van

- **Action**
  - Walking
  - Speaking
  - Running
  - Sitting down
  - Standing
  - Singing
  - Handshaking
  - Swimming
  - Throwing
  - Greeting

- **Scene**
  - Court
  - Urban
  - Kitchen
  - Hospital
  - Highway
  - Bakery
  - Flood
  - Field
  - Desert
  - Beach

- **People**
  - Groom
  - Researcher
  - Indian person
  - Two people
  - Teenager
  - Politician
  - Athlete
  - Baby
  - Adult male
  - Adult female

- **Animal**
  - Flamingo
  - Scorpion
  - Koala
  - Horse
  - Wild animal
  - Insect
  - Dolphin
  - Cow
  - Cat
  - Bird

- **Attribute**
  - Triangle
  - Professional Video
  - Cartoon
  - Still image
  - Scene text
  - Overlaid text
  - Moon light
  - Junk frame
  - Graphic
  - Amateur Video
Q1: How many concepts?

- More is better, especially for small vocabularies
- Include at least 200 concepts
Q2: What concept types?

- Derive the vocabulary concepts
  - Only from a particular concept type? (Single type)
  - From various concept types? (Joint type)

<table>
<thead>
<tr>
<th>Scene (128)</th>
<th>Single</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.142</td>
<td>0.168</td>
</tr>
</tbody>
</table>
Q2: What concept types?

- Derive the vocabulary concepts
  - Only from a particular concept type? (Single type)
  - From various concept types? (Joint type)

<table>
<thead>
<tr>
<th>Vocab.</th>
<th>Object (670)</th>
<th>Action (34)</th>
<th>Scene (128)</th>
<th>People (78)</th>
<th>Animal (321)</th>
<th>Attribute (45)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.259</td>
<td><strong>0.279</strong></td>
<td>0.067</td>
<td><strong>0.076</strong></td>
<td>0.142</td>
<td><strong>0.168</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joint</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MAP values for different concept types.
Q3: How Accurate?

- Decrease concept detector accuracies to observe how event detection performance respond.
Q3: How Accurate?

- Decrease concept detector accuracies to observe how event detection performance respond.
Recommendations

- For event detection using concept vocabularies
  - Include at least 200 detectors
  - Diversity of concept types is important
  - Concept detector accuracy is not critical
Backup Slides
<table>
<thead>
<tr>
<th>Event</th>
<th>Animal (321)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single</td>
<td>Joint</td>
</tr>
<tr>
<td>Attempting board trick</td>
<td>0.120</td>
<td>0.271</td>
</tr>
<tr>
<td>Feeding animal</td>
<td><strong>0.073</strong></td>
<td>0.045</td>
</tr>
<tr>
<td>Landing fish</td>
<td>0.323</td>
<td>0.36</td>
</tr>
<tr>
<td>Wedding ceremony</td>
<td>0.162</td>
<td>0.388</td>
</tr>
<tr>
<td>Working wood working project</td>
<td>0.116</td>
<td>0.167</td>
</tr>
<tr>
<td>Birthday party</td>
<td>0.139</td>
<td>0.239</td>
</tr>
<tr>
<td>Changing vehicle tire</td>
<td>0.054</td>
<td>0.153</td>
</tr>
<tr>
<td>Flash mob gathering</td>
<td>0.415</td>
<td>0.475</td>
</tr>
<tr>
<td>Getting vehicle unstuck</td>
<td>0.294</td>
<td>0.338</td>
</tr>
<tr>
<td>Grooming animal</td>
<td><strong>0.146</strong></td>
<td>0.127</td>
</tr>
<tr>
<td>Making sandwich</td>
<td>0.07</td>
<td>0.176</td>
</tr>
<tr>
<td>Parade</td>
<td>0.126</td>
<td>0.275</td>
</tr>
<tr>
<td>Parkour</td>
<td>0.089</td>
<td>0.356</td>
</tr>
<tr>
<td>Repairing appliance</td>
<td>0.104</td>
<td>0.259</td>
</tr>
<tr>
<td>Working sewing project</td>
<td>0.194</td>
<td>0.238</td>
</tr>
<tr>
<td>Attempting bike trick</td>
<td>0.129</td>
<td>0.392</td>
</tr>
<tr>
<td>Cleaning appliance</td>
<td>0.029</td>
<td>0.058</td>
</tr>
<tr>
<td>Dog show</td>
<td><strong>0.555</strong></td>
<td>0.512</td>
</tr>
<tr>
<td>Giving directions location</td>
<td>0.016</td>
<td>0.029</td>
</tr>
<tr>
<td>Marriage proposal</td>
<td>0.018</td>
<td>0.05</td>
</tr>
<tr>
<td>Renovating home</td>
<td>0.085</td>
<td>0.192</td>
</tr>
<tr>
<td>Rock climbing</td>
<td>0.309</td>
<td>0.322</td>
</tr>
<tr>
<td>Town hall meeting</td>
<td>0.266</td>
<td>0.379</td>
</tr>
<tr>
<td>Winning race without vehicle</td>
<td>0.088</td>
<td>0.138</td>
</tr>
<tr>
<td>Working metal crafts project</td>
<td>0.019</td>
<td>0.038</td>
</tr>
<tr>
<td>Event</td>
<td>Specific</td>
<td>General</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>----------</td>
<td>---------</td>
</tr>
<tr>
<td>Attempting board trick</td>
<td>0.090</td>
<td>0.108</td>
</tr>
<tr>
<td>Feeding animal</td>
<td>0.041</td>
<td>0.042</td>
</tr>
<tr>
<td>Landing fish</td>
<td>0.113</td>
<td>0.107</td>
</tr>
<tr>
<td>Wedding ceremony</td>
<td>0.071</td>
<td>0.14</td>
</tr>
<tr>
<td>Working wood working project</td>
<td>0.083</td>
<td>0.065</td>
</tr>
<tr>
<td>Birthday party</td>
<td>0.078</td>
<td>0.135</td>
</tr>
<tr>
<td>Changing vehicle tire</td>
<td>0.058</td>
<td>0.062</td>
</tr>
<tr>
<td>Flash mob gathering</td>
<td>0.301</td>
<td>0.284</td>
</tr>
<tr>
<td>Getting vehicle unstuck</td>
<td>0.195</td>
<td>0.246</td>
</tr>
<tr>
<td>Grooming animal</td>
<td>0.064</td>
<td>0.079</td>
</tr>
<tr>
<td>Making sandwich</td>
<td>0.059</td>
<td>0.089</td>
</tr>
<tr>
<td>Parade</td>
<td>0.073</td>
<td>0.203</td>
</tr>
<tr>
<td>Parkour</td>
<td>0.104</td>
<td>0.226</td>
</tr>
<tr>
<td>Repairing appliance</td>
<td>0.111</td>
<td>0.098</td>
</tr>
<tr>
<td>Working sewing project</td>
<td>0.076</td>
<td>0.075</td>
</tr>
<tr>
<td>Attempting bike trick</td>
<td>0.044</td>
<td>0.08</td>
</tr>
<tr>
<td>Cleaning appliance</td>
<td>0.125</td>
<td>0.092</td>
</tr>
<tr>
<td>Dog show</td>
<td>0.219</td>
<td>0.178</td>
</tr>
<tr>
<td>Giving directions location</td>
<td>0.028</td>
<td>0.019</td>
</tr>
<tr>
<td>Marriage proposal</td>
<td>0.013</td>
<td>0.017</td>
</tr>
<tr>
<td>Renovating home</td>
<td>0.023</td>
<td>0.074</td>
</tr>
<tr>
<td>Rock climbing</td>
<td>0.178</td>
<td>0.156</td>
</tr>
<tr>
<td>Town hall meeting</td>
<td>0.064</td>
<td>0.226</td>
</tr>
<tr>
<td>Winning race without vehicle</td>
<td>0.102</td>
<td>0.102</td>
</tr>
<tr>
<td>Working metal crafts project</td>
<td>0.040</td>
<td>0.021</td>
</tr>
</tbody>
</table>
How Accurate?

☐ Add noise to detector predictions

Add noise to detector predictions

SVM Prediction Score | Frequency
--- | ---
-2 | 500
-1.5 | 1000
-1 | 1500
-0.5 | 2000
0 | 2500
0.5 | 500
1 | 1000
1.5 | 1500
2 | 2000
2.5 | 2500

Confident Area

Confident Area
Concept Correlations

![Correlation Matrix]

- **Object**
- **Action**
- **Scene**
- **People**
- **Animal**
- **Attribute**

- **No correlation**
- **High Correlation**

The heat map visualizes the correlations between different concepts, with colors ranging from blue (low correlation) to red (high correlation). The diagonal line indicates no correlation.
34: How Accurate?

- Decrease concept detector accuracies to observe how event detection performance respond

- Approach 1: Train less sophisticated detectors
  - 100%-3SIFT-SP
  - 30%-3SIFT-SP
  - 30%-SIFT-SP
  - 30%-SIFT

- Approach 2: Impose noise into detector predictions
Q4: How Accurate?

- Results for vocabulary of 346 concepts
- Relatively robust against concept detector accuracies

<table>
<thead>
<tr>
<th>Detectors</th>
<th>100%-SIFT-SP</th>
<th>30%-SIFT-SP</th>
<th>30%-SIFT-SP</th>
<th>30%-SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.206</td>
<td>0.189</td>
<td>0.182</td>
<td>0.185</td>
</tr>
</tbody>
</table>

MED dataset