Events

Cees Snoek

University of Amsterdam
The Netherlands
Motivation: Internet of things that video
Technology: self-driving cars
Forensics: Analyzing terrorist behavior
Well-being: elderly monitoring

Figure 1. Examples of interaction patterns in a nursing home

Chen et al. MM 2004
Studentencomplexen krijgen meer veiligheidscamera’s

Studentencomplexen zijn niet onveiliger dan andere woonwijken. Wel moeten in verband met incidenten, waarbij studentes zijn aangevallen, veiligheidsmaatregelen getroffen worden.
What is an event?

News events: *earthquake, abdication, product launch*

Sport events: *scoring goal, ace serve, slam dunk*

Social events: *concert, debates, exhibitions*

Every day events: interactions of people and objects

- Repairing an appliance
- Working on sewing project
- Grooming an animal
- Birthday party
Goal

Board trick

Assembling a shelter

Birthday party

Recognize and explain event as it happens in video
This lecture

We study event recognition

I. Data, data, data

II. Event classification

III. Event retrieval
Prelude

DATA, DATA, DATA
The early years 1995-2010

Progress was slow

– Lack of data
– Lack of benchmarks
– Lack of community
– Lack of urgency
Goalgle: 9 hrs of test video...
CCV: Columbia Consumer Video Database

http://www.ee.columbia.edu/ln/dvmm/CCV/
CCV snapshot

- **# videos**: 9,317  
  - (210 hrs in total)
- **video genre**  
  - unedited consumer videos
- **video source**  
  - YouTube.com
- **average length**  
  - 80 seconds
- **# defined categories**  
  - 20
- **annotation method**  
  - Amazon Mechanical Turk

![Bar chart showing the number of positive videos per category.](chart)
TRECVID benchmark

International competition

Promote progress in video retrieval research

Open data, tasks, evaluation and innovation
## Internet video collections

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Designated Uses</th>
<th>Target sizes</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot</td>
<td>2010</td>
<td>Development collection Test collection</td>
<td>1,723 clips 1,742 clips (100 hours)</td>
</tr>
<tr>
<td>Development (DEV)</td>
<td>2011</td>
<td>Split into two subsets: (1) Transparent (DEV-T) (2) Opaque (DEV-O)</td>
<td>44K clips, (~1400 hours)</td>
</tr>
<tr>
<td></td>
<td>2012-2015</td>
<td>(1) and (2) merged to a single training collection</td>
<td></td>
</tr>
<tr>
<td>Progress</td>
<td>2012-2015: test collection</td>
<td>120K clips, 4000 hrs</td>
<td>No clip content annotation</td>
</tr>
<tr>
<td>Novel 1</td>
<td>2014: test collection</td>
<td>120K clips, 4000 hrs</td>
<td>No clip content annotation</td>
</tr>
<tr>
<td>Novel 2</td>
<td>2015: test collection</td>
<td>120K clips, 4000 hrs</td>
<td>No clip content annotation</td>
</tr>
</tbody>
</table>
The TRECVID MED ‘11 events

**Process-Observed Events**
- Attempting a board trick
- Feeding an animal
- Landing a fish
- Working on a woodworking project

**Life Events**
- Wedding ceremony

**Process-Observed Events**
- Changing a vehicle tire
- Getting a vehicle unstuck
- Grooming an animal
- Making a sandwich
- Parkour
- Repairing an appliance
- Working on a sewing project

**Life Events**
- Birthday party
- Flash mob gathering
- Parade
Example Event Kit

Event Name:
Working on a woodworking project

Definition:
One or more people fashion an object out of wood.

Event Explication:
Woodworking is a popular hobby that involves crafting an object out of wood. Typical woodworking projects may range from creating large pieces of furniture to small decorative items or toys. The process for making objects out of wood can include cutting wood into smaller pieces .... (continues)

Evidential Description:

scene: Often indoors in a workshop, garage, artificial lighting.
Occasionally outdoors

objects/people: Woodworking tools (automatic or non-automatic saws, sander, knife), paint, stains, sawhorses, toolbox, safety goggles

activities: Cutting and shaping wood, attaching pieces of wood together, smoothing/sanding wood

audio: power tool sounds; hand tool sounds (hammer, saw, etc.); narration of process

Exemplars:
HVC334271.mp4, HVC393428.mp4, HVC875424.mp4, etc.

Target User:
An *Internet information analyst* or *experienced Internet searcher* with event-specialized knowledge.
Part I

CLASSIFICATION
Chapter 1

FEATURE ENCODING

Several slides by: Yu-Gang Jiang
Solution 1: Feature encoding

Represent video as low-level feature vector

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

\textbf{Good recognition accuracy, no interpretation}

Y.G. Jiang et al. TRECVID10
P. Natarjan et al., CVPR12
Wang et al., ICCV13
\textit{and many others}
Winner TRECVID 2010

Contribution per modality

More is better, feature fusion strong fundamental
Audiovisual features

- **SIFT (visual)**
  - D. Lowe, IJCV 04.

- **STIP (visual)**
  - I. Laptev, IJCV 05.

- **MFCC (audio)**
Bag-of-X representation

- $X = \text{SIFT} / \text{STIP} / \text{MFCC}$
- **Soft weighting** (Jiang, Ngo and Yang, ACM CIVR 2007)
Results

- Measured by Average Precision (AP)

<table>
<thead>
<tr>
<th></th>
<th>Assembling a shelter</th>
<th>Batting a run in</th>
<th>Making a cake</th>
<th>Mean AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual STIP</td>
<td>0.468</td>
<td>0.719</td>
<td>0.476</td>
<td>0.554</td>
</tr>
<tr>
<td>Visual SIFT</td>
<td>0.353</td>
<td>0.787</td>
<td>0.396</td>
<td>0.512</td>
</tr>
<tr>
<td>Audio MFCC</td>
<td>0.249</td>
<td>0.692</td>
<td>0.270</td>
<td>0.404</td>
</tr>
<tr>
<td>STIP+SIFT</td>
<td>0.508</td>
<td>0.796</td>
<td>0.476</td>
<td>0.593</td>
</tr>
<tr>
<td>STIP+SIFT+MFCC</td>
<td><strong>0.533</strong></td>
<td><strong>0.873</strong></td>
<td><strong>0.493</strong></td>
<td><strong>0.633</strong></td>
</tr>
</tbody>
</table>

- STIP works the best for event detection
- The 3 features are highly complementary!
2011 event detection results

Easy:
Flash mob

Hard:
Grooming an animal

All systems rely predominantly on bag-of-features, no notion whether event really happened.
2012 & 2013 winner: Inria LEAR

Low-level features
- SIFT
- Color
- MFCC
- Improved trajectories

Encoding
- Spatial Fisher vector
- Spatial Fisher vector
- Fisher vector
- Fisher vector

High-level features
- OCR
- ASR
- Bag-of-words
- Bag-of-words

Classifier

Classification
2014 winner: CMU

Winning system combined many multimedia features, with huge computation budget, deep learning key?
2015 winner: ImageNet-Shuffle - UvA

Leverage complete, but reorganized ImageNet for pre-training
Outperform standard networks, maintain benefits of fusion

http://tinyurl.com/imagenetshuffle
Conclusion on feature encodings

• The combination of audio-visual features is key for good video event recognition
  – MBH + Fisher vector best single feature
  – Best single feature from deep convolutional nets

• Many start to explore temporal deep learning
  – 3D convolutions
  – Recurrent neural networks
  – ...

Good recognition accuracy, limited interpretation
Chapter 2

SEMANTIC ENCODING

Joint work with Amirhossein Habibian & Masoud Mazloom
Solution 2: Semantic encoding

Represent video as concept score histogram

– Detectors from deep learning, Fisher vectors, etc.
– Annotated examples from ImageNet, Flickr, etc.

Vocabulary for semantic encoding mostly driven by ad hoc concept detector availability.

Naphade et al. TMM02
Ebadollahi et al., ICME06
Snoek et al., PAMI06
Gkalelis et al., CBMI11
Merler et al., TMM12
and many others
Semantic encodings for video

1. How many concepts?
2. What concept types?
3. Which concepts?
4. How accurate?
5. How to select?
Experimental setup

**MED**: TRECVID Multimedia Event Detection 2012
13,274 videos (66% train, 34% test)
25 event categories, *marriage proposal, grooming animal, etc.*

**CCV**: Columbia Consumer Video
9,317 videos (50% train, 50% test)
15 event categories, *music performance, graduation, etc.*

Vocabulary sampled from 1,346 concept detectors
Annotations by ImageNet Challenge11 and TRECVID SIN12
Color Fisher coding with spatial pooling and linear SVM
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
1. How many concepts?

More is better, but include at least 200

Mean Average Precision

Vocabulary Size

MED dataset

CCV dataset

More is better, but include at least 200
2. What concept types?

Derive the vocabulary concepts

*Single*: Only from a particular concept type?

*Joint*: From various concept types?

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

<table>
<thead>
<tr>
<th>Vocab.</th>
<th>Single</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.259</td>
<td>0.279</td>
</tr>
<tr>
<td></td>
<td>0.067</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>0.142</td>
<td>0.168</td>
</tr>
<tr>
<td></td>
<td>0.082</td>
<td>0.123</td>
</tr>
<tr>
<td></td>
<td>0.158</td>
<td>0.239</td>
</tr>
<tr>
<td></td>
<td>0.063</td>
<td>0.082</td>
</tr>
</tbody>
</table>

Small difference

Big difference

<table>
<thead>
<tr>
<th>Vocab.</th>
<th>Single</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.307</td>
<td>0.335</td>
</tr>
<tr>
<td></td>
<td>0.197</td>
<td>0.217</td>
</tr>
<tr>
<td></td>
<td>0.249</td>
<td>0.285</td>
</tr>
<tr>
<td></td>
<td>0.229</td>
<td>0.265</td>
</tr>
<tr>
<td></td>
<td>0.265</td>
<td>0.310</td>
</tr>
<tr>
<td></td>
<td>0.178</td>
<td>0.220</td>
</tr>
</tbody>
</table>

In general, a diverse vocabulary is better
<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>
</tbody>
</table>
Concept correlations

![Heatmap showing correlations between different concept categories: Object, Action, Scene, People, Animal, Attribute. The color bar indicates high correlation (1) and no correlation (0). The circled area highlights a region of high correlation.](image)

*Plotted for MED dataset*
Semantic encodings for video

1. How many concepts?
2. What concept types?
3. Which concepts?
4. How accurate?
5. How to select?
3. Which concepts?

General/specific concepts are identified manually

General: human, vegetation, outdoor etc.
Specific: salmon, cheese, sand castle etc.

Derive the vocabulary concepts

Only from specific concepts?
Only from general concepts?
Mixture of specific and general concepts?
3. Which concepts?

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Specific</th>
<th>General</th>
<th>Mixture</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MED dataset</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP</td>
<td>0.094</td>
<td>0.117</td>
<td>0.130</td>
</tr>
<tr>
<td><strong>CCV dataset</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAP</td>
<td>0.208</td>
<td>0.232</td>
<td>0.260</td>
</tr>
</tbody>
</table>

Specific and general concepts should be mixed
<table>
<thead>
<tr>
<th>Event</th>
<th>Specific</th>
<th>General</th>
<th>Mixture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attempting board trick</td>
<td>0.090</td>
<td>0.108</td>
<td>0.130</td>
</tr>
<tr>
<td>Feeding animal</td>
<td>0.041</td>
<td>0.042</td>
<td>0.045</td>
</tr>
<tr>
<td>Landing fish</td>
<td>0.113</td>
<td>0.107</td>
<td>0.139</td>
</tr>
<tr>
<td>Wedding ceremony</td>
<td>0.071</td>
<td>0.14</td>
<td>0.164</td>
</tr>
<tr>
<td>Working wood working project</td>
<td>0.083</td>
<td>0.065</td>
<td>0.073</td>
</tr>
<tr>
<td>Birthday party</td>
<td>0.078</td>
<td>0.135</td>
<td>0.138</td>
</tr>
<tr>
<td>Changing vehicle tire</td>
<td>0.058</td>
<td>0.062</td>
<td>0.071</td>
</tr>
<tr>
<td>Flash mob gathering</td>
<td>0.301</td>
<td>0.284</td>
<td>0.337</td>
</tr>
<tr>
<td>Getting vehicle unstuck</td>
<td>0.195</td>
<td>0.246</td>
<td>0.282</td>
</tr>
<tr>
<td>Grooming animal</td>
<td>0.064</td>
<td>0.079</td>
<td>0.081</td>
</tr>
<tr>
<td>Making sandwich</td>
<td>0.059</td>
<td>0.089</td>
<td>0.119</td>
</tr>
<tr>
<td>Parade</td>
<td>0.073</td>
<td>0.203</td>
<td>0.161</td>
</tr>
<tr>
<td>Parkour</td>
<td>0.104</td>
<td>0.226</td>
<td>0.210</td>
</tr>
<tr>
<td>Repairing appliance</td>
<td>0.111</td>
<td>0.098</td>
<td>0.101</td>
</tr>
<tr>
<td>Working sewing project</td>
<td>0.076</td>
<td>0.075</td>
<td>0.082</td>
</tr>
<tr>
<td>Attempting bike trick</td>
<td>0.044</td>
<td>0.08</td>
<td>0.09</td>
</tr>
<tr>
<td>Cleaning appliance</td>
<td>0.125</td>
<td>0.092</td>
<td>0.123</td>
</tr>
<tr>
<td>Dog show</td>
<td>0.219</td>
<td>0.178</td>
<td>0.23</td>
</tr>
<tr>
<td>Giving directions location</td>
<td>0.028</td>
<td>0.019</td>
<td>0.053</td>
</tr>
<tr>
<td>Marriage proposal</td>
<td>0.013</td>
<td>0.017</td>
<td>0.025</td>
</tr>
<tr>
<td>Renovating home</td>
<td>0.023</td>
<td>0.074</td>
<td>0.083</td>
</tr>
<tr>
<td>Rock climbing</td>
<td>0.178</td>
<td>0.156</td>
<td>0.194</td>
</tr>
<tr>
<td>Town hall meeting</td>
<td>0.064</td>
<td>0.226</td>
<td>0.158</td>
</tr>
<tr>
<td>Winning race without vehicle</td>
<td>0.102</td>
<td>0.102</td>
<td>0.117</td>
</tr>
<tr>
<td>Working metal crafts project</td>
<td>0.040</td>
<td>0.021</td>
<td>0.036</td>
</tr>
</tbody>
</table>
4. How accurate?

How important is the concept detector accuracy?

Decrease concept detector accuracies to observe how event detection performance responds.

Approach: Train less sophisticated detectors
Approach: Four detector settings

All examples / ColorSIFT / Spatial Pyramids

30% of examples / ColorSIFT / Spatial Pyramids

30% of examples / SIFT / Spatial Pyramids

30% of examples / SIFT
Train less sophisticated detectors

### MED dataset

<table>
<thead>
<tr>
<th>Detectors</th>
<th>100% Examples</th>
<th>30% Examples</th>
<th>30% Examples</th>
<th>30% Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ColorSIFT</td>
<td>ColorSIFT</td>
<td>SIFT</td>
<td>SIFT</td>
</tr>
<tr>
<td></td>
<td>Spatial</td>
<td>Spatial</td>
<td>Spatial</td>
<td>Spatial</td>
</tr>
<tr>
<td></td>
<td>Pyramid</td>
<td>Pyramid</td>
<td>Pyramid</td>
<td>Pyramid</td>
</tr>
<tr>
<td>MAP</td>
<td>0.206</td>
<td>0.189</td>
<td>0.182</td>
<td>0.185</td>
</tr>
</tbody>
</table>

### CCV dataset

<table>
<thead>
<tr>
<th>Detectors</th>
<th>100% Examples</th>
<th>30% Examples</th>
<th>30% Examples</th>
<th>30% Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ColorSIFT</td>
<td>ColorSIFT</td>
<td>SIFT</td>
<td>SIFT</td>
</tr>
<tr>
<td></td>
<td>Spatial</td>
<td>Spatial</td>
<td>Spatial</td>
<td>Spatial</td>
</tr>
<tr>
<td></td>
<td>Pyramid</td>
<td>Pyramid</td>
<td>Pyramid</td>
<td>Pyramid</td>
</tr>
<tr>
<td>MAP</td>
<td>0.359</td>
<td>0.371</td>
<td>0.354</td>
<td>0.353</td>
</tr>
</tbody>
</table>

More sophisticated detectors have only minor influence on the overall event recognition accuracy.
Semantic encodings for video

1. How many concepts?
2. What concept types?
3. Which concepts?
4. How accurate?
5. How to select?
5. Motivation

More is better, but include at least 200
Example for: *Landing a fish in*

A vocabulary of 400 concepts is more accurate than using all
Example for: *Wedding ceremony*

A vocabulary of 500 concepts is more accurate than using all
Value of individual concepts

<table>
<thead>
<tr>
<th>Board trick</th>
<th>Concept</th>
<th>AP</th>
<th>Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skating</td>
<td>0.194</td>
<td>1,300</td>
<td></td>
</tr>
<tr>
<td>Road</td>
<td>0.171</td>
<td>1,096</td>
<td></td>
</tr>
<tr>
<td>Snow</td>
<td>0.162</td>
<td>1,013</td>
<td></td>
</tr>
<tr>
<td>Snowplow</td>
<td>0.123</td>
<td>540</td>
<td></td>
</tr>
<tr>
<td>Ski</td>
<td>0.119</td>
<td>1,096</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wedding ceremony</th>
<th>Concept</th>
<th>AP</th>
<th>Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Church</td>
<td>0.396</td>
<td>1,300</td>
<td></td>
</tr>
<tr>
<td>Altar</td>
<td>0.324</td>
<td>1,300</td>
<td></td>
</tr>
<tr>
<td>Gown</td>
<td>0.306</td>
<td>1300</td>
<td></td>
</tr>
<tr>
<td>Groom</td>
<td>0.288</td>
<td>1,280</td>
<td></td>
</tr>
<tr>
<td>Suit</td>
<td>0.251</td>
<td>1,300</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Flash mob gathering</th>
<th>Concept</th>
<th>AP</th>
<th>Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crowd</td>
<td>0.280</td>
<td>2,341</td>
<td></td>
</tr>
<tr>
<td>3 or more people</td>
<td>0.214</td>
<td>2,099</td>
<td></td>
</tr>
<tr>
<td>People marching</td>
<td>0.205</td>
<td>624</td>
<td></td>
</tr>
<tr>
<td>Street battle</td>
<td>0.202</td>
<td>1,300</td>
<td></td>
</tr>
<tr>
<td>Meeting</td>
<td>0.186</td>
<td>340</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Basketball</th>
<th>Concept</th>
<th>AP</th>
<th>Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basketball</td>
<td>0.488</td>
<td>1,300</td>
<td></td>
</tr>
<tr>
<td>Throw ball</td>
<td>0.485</td>
<td>811</td>
<td></td>
</tr>
<tr>
<td>Throwing</td>
<td>0.432</td>
<td>1,300</td>
<td></td>
</tr>
<tr>
<td>Indoor sport venue</td>
<td>0.355</td>
<td>1,300</td>
<td></td>
</tr>
<tr>
<td>Gym</td>
<td>0.337</td>
<td>153</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Swimming</th>
<th>Concept</th>
<th>AP</th>
<th>Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swimming</td>
<td>0.698</td>
<td>1,300</td>
<td></td>
</tr>
<tr>
<td>Swimming pool</td>
<td>0.621</td>
<td>1,300</td>
<td></td>
</tr>
<tr>
<td>Underwater</td>
<td>0.432</td>
<td>1,300</td>
<td></td>
</tr>
<tr>
<td>Stingray</td>
<td>0.227</td>
<td>1,300</td>
<td></td>
</tr>
<tr>
<td>Waterscape/Waterfront</td>
<td>0.211</td>
<td>604</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parade</th>
<th>Concept</th>
<th>AP</th>
<th>Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>People marching</td>
<td>0.318</td>
<td>624</td>
<td></td>
</tr>
<tr>
<td>Urban scenes</td>
<td>0.155</td>
<td>1,403</td>
<td></td>
</tr>
<tr>
<td>Police van</td>
<td>0.150</td>
<td>1,300</td>
<td></td>
</tr>
<tr>
<td>3 or more people</td>
<td>0.138</td>
<td>2,099</td>
<td></td>
</tr>
<tr>
<td>Streets</td>
<td>0.135</td>
<td>1,300</td>
<td></td>
</tr>
</tbody>
</table>

Note the semantic correspondence between good performing concepts and events
Research question 5.

Is it possible to learn the semantic encoding of an event from examples?
Idea

Formalize subset selection as importance sampling

Cross-entropy optimization
1. Sample semantic subset
2. Evaluate semantic subset
3. Update sampling parameters

Near-optimal solution

Mazloom et al., TMM 2014

Concept vocabulary
All concepts (•) vs selected concepts (*)

Encoding based on selected concepts always better
Concept subsets are descriptive

Font size correlates with importance
Failure case

Why is an ‘Abacus’ descriptive for Birthday?

Example training examples for candle and abacus
Recommendations

For event recognition using semantic encodings

1. Include at least 200 detectors
2. Diversity of concept types is important
3. Both the general and specific concepts are required
4. Concept detector accuracy is not critical
5. A descriptive concept subset can be learned from examples

Part II

RETRIEVAL

Joint work with Amirhossein Habibian & Masoud Mazloom
Hypothesis

As events become more and more specific, it is unrealistic to assume that ample examples to learn from will be commonly available.
Goal

Event Name: Winning a race without a vehicle
Definition: An individual (or more) succeeds in reaching a pre-determined destination before all other individuals, without vehicle assistance or assistance of a horse or other animal. Racing generally involves accomplishing a task in less time than other competitors. The only type of racing considered relevant for the purposes of this event is the type where the task is traveling to a destination, completed by a person(s) without assistance of a vehicle or animal. Different types of races involve different types of human ...
Feature embedding fails

Representing videos as histograms of low-level features

Problem: demands examples

[Jiang et al., TRECVID 2010] [Natarajan et al., CVPR 2012] [Wang et al., ICCV 2013] and many others
Solution

The key to event recognition when examples are absent is to have a **lingual** video representation.

Once the video is represented in a textual form, standard retrieval metrics can be used.
Event recognition, without examples

This talk Part II
This part: three lingual representations

Concept embedding

Tag embedding

Video2vec embedding
Chapter 3

CONCEPT EMBEDDING
Representing videos as histograms of concept scores

Problem: define, annotate and train concept classifiers

Deep convolutional neural network

Local descriptors
- Visual descriptors
  - SIFT, HOG, GIST, …
- Video descriptors
  - MBH, STIP, …
- Audio descriptors
  - MFCC, AIM, …

Feature embedding
- Bag-of-words
- VLAD
- Fisher vector
- Audio-visual BoW

Classification
- Attribute detection
- Concept detection

[Einadola et al., ICME 2006] [Merler et al., TMM 2012] [Habibian et al., CVIU 2014] and many others
Label composition trick

Expanding the labels by logical operations

• AND, OR, ...

Habibian et al. ICMR 2014
Label composition trick

Expanding the labels by logical operations

• AND, OR, ...

<table>
<thead>
<tr>
<th>Ride</th>
<th>Motorcycle</th>
<th>Bike</th>
<th>Bike-AND-Ride</th>
<th>Bike-OR-Motorcycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Concept Annotations
Motivation

Expanding the vocabulary for *free*

Composite concepts can be easier to detect
- boat-AND-sea
- bear-AND-cage
- man-OR-woman

Composite concepts can be more indicative of the event
- bike-AND-ride for *attempting a bike trick*
Learning composite concepts

For a vocabulary of n concepts, there are $B_n$ disjoint compositions

- Bell number: $B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k$
- Not all of them are useful

Which concepts should be composed together?

- NP-hard problem, equivalent to set-partitioning
- Approximated by a greedy search algorithm
Qualitative results

Top ranked videos for *flash mob gathering*

Most dominant concepts in the video representation

<table>
<thead>
<tr>
<th>Detected Videos</th>
<th>Composite Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Image 1]</td>
<td>Group-AND-Dance-AND-Shopping</td>
</tr>
<tr>
<td>![Image 2]</td>
<td>Group-AND-Dance-AND-Shopping</td>
</tr>
</tbody>
</table>
Composite concepts

Label composition leads to a more comprehensive concept embedding

Still need to define, annotate and train concept classifiers

Greedy search algorithm slow
Drawbacks of concept discovery

Big computational effort

Many concepts are rare, insufficient examples to train reliable visual classifiers

Selection is based on visual prediction accuracy only, descriptiveness is ignored

Contextual information is lost, since concepts are learned independently by binary classifiers.
Chapter 4

TAG EMBEDDING

Masoud Mazloom, Xirong Li, and Cees G. M. Snoek,
TagBook: A Semantic Video Representation without Supervision for Event Detection,
IEEE Transactions on Multimedia, in press.
Idea

Embedding based on freely available social tagged videos only

Without the need for training any intermediate concept detectors
Inspiration

Xirong Li et al, TMM 2009
TagBook: embedding derived from social tags

Social-tagged web videos

<table>
<thead>
<tr>
<th>Video data</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Video data" /></td>
<td>woman, outdoor, metal-crafts-project, welding machine</td>
</tr>
<tr>
<td><img src="image2" alt="Video data" /></td>
<td>man, kitchen, metallic, cleaning, oven, spray, glasses,</td>
</tr>
<tr>
<td><img src="image3" alt="Video data" /></td>
<td>man, snowboard, snow, board-trick,</td>
</tr>
<tr>
<td><img src="image4" alt="Video data" /></td>
<td>man, climb-on, wall, gym, rock-climbing</td>
</tr>
</tbody>
</table>

TagBook = {woman, outdoor, metal-crafts-project, welding machine, man, kitchen, ..., wall, gym, rock-climbing}
It is advantageous to select most frequent tags in TagBook
Chapter 5

VIDEO2VEC EMBEDDING

Research question

Can we **learn the embedding** from videos and their stories?

Story usually highlights the key concepts in video

Videos and stories are freely available, *i.e.* YouTube
Multimedia embeddings

Joint space where $x_i \cdot W \approx y_i \cdot A$

Explicitly relate training $W$ and $A$ from multimedia

$W = \text{Visual projection matrix}$  individual term classifiers
$A = \text{Textual projection matrix}$  select/group terms

[Rasiwasa et al., MM 2010] [Weston et al., IJCAI 2011] [Akata et al., CVPR 2013] [Das et al., WSDM 2013]
Video2vec: Embed the story of a video

Design criteria: learn $W$ and $A$ such that

*Descriptiveness:* preserve video descriptions

*Predictability:* recognize terms from video content
Key observation: Compelling forces

Crazy guy doing insane stunts on bike
Why is this important?

Grouping terms:
  Number of classes is reduced

Training classifiers per group:
  More positive examples available per group

We can train from freely available web data
Key contribution: Joint optimization

Jointly optimize for descriptiveness and predictability

\[ L_{VS}(A, W) = \min_{S} L_d(A, S) + L_p(S, W) \]

Hyperparameter: size of the embedding \( S \)
- \( L_d \) Loss function for descriptiveness
- \( L_p \) Loss function for predictability

Video2vec connects the two loss functions
Video2vec objectives: **descriptiveness**

**Objective 1:** The Video2vec embedding should be **descriptive**

\[
L_d(A, S) = \frac{1}{N} \sum_{i=1}^{N} \| y_i - A s_i \|_2^2 + \lambda_a \Omega(A) + \lambda_s \Psi(S)
\]

- **Original transcriptions**
- **Reconstructed terms**
- **Regularizers**

Essentially latent semantic indexing with L2 rather than an L1 norm
Video2vec objectives: **predictability**

**Objective 2:** The Video2vec embedding should be *predictable*

\[
L_p(S, W) = \frac{1}{N} \sum_{i=1}^{N} \| s_i - W^\top x_i \|_2^2 + \lambda_w \Theta(W)
\]

- **Video2vec embedding**
- **Video feature embedding**
- **Regularizer**
Video2vec: Training

Set of videos and their captions

Encode video features $x_i$
Any feature (combination) will do

Encode video descriptions $y_i$
Bag-of-words of terms
VideoStory46K dataset

Videos and title descriptions from YouTube

46K videos, 19K unique terms in descriptions

Seeded from video event descriptions

Filters to remove low quality videos

Available for download: www.mediamill.nl
Video2vec: Training (2)

Using *Stochastic Gradient Descent*:

Choose random sample

Compute sample gradient wrt objective

\[
\begin{align*}
\nabla_A L_{VS} &= -2 (y_t - As_t) s_t^T + \lambda_a A, \\
\nabla_W L_{VS} &= -2 x_t (s_t - W^T x_t)^T + \lambda_w W, \text{ and} \\
\nabla_{s_t} L_{VS} &= 2 \left[ s_t - W^T x_t - A^T (y_t - As_t) \right] + \lambda_s s_t.
\end{align*}
\]

Update parameters with step-size \( \eta \)
Video2vec at work

1. Project visual features
   \[ s_i = W^T x_i, \]

2. Translate to text
   \[ \hat{y}_i = A s_i, \]

3. Cosine distance for matching
   \[ s_e(x_i) = \frac{y^e^T \hat{y}^e_i}{||y^e|| \cdot ||\hat{y}^e||} \]
Video2vec predicted terms

- *non-motorized vehicle repair*
  - bicycle
  - repair
  - tire
  - wheel
  - chain

- *horse riding competition*
  - race
  - run
  - ride
  - jump
  - horse

- *renovating a home*
  - tile
  - window
  - wall
  - garage
  - install
# State-of-the-art event retrieval

<table>
<thead>
<tr>
<th>Authors</th>
<th>Published</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Habibian et al.</td>
<td>ICMR 2014</td>
<td>6.4</td>
</tr>
<tr>
<td>Ye et al.</td>
<td>MM 2015</td>
<td>9.0</td>
</tr>
<tr>
<td>Chang et al.</td>
<td>IJCAI 2015</td>
<td>9.6</td>
</tr>
<tr>
<td>Mazloom et al.</td>
<td>ICMR 2015</td>
<td>11.9</td>
</tr>
<tr>
<td>Wu et al.</td>
<td>CVPR 2014</td>
<td>12.7</td>
</tr>
<tr>
<td>Jiang et al.</td>
<td>AAAI 2015</td>
<td>12.9</td>
</tr>
<tr>
<td>Mazloom et al.</td>
<td>TMM 2016</td>
<td>12.9</td>
</tr>
<tr>
<td>Liang et al.</td>
<td>MM 2015</td>
<td>18.3</td>
</tr>
<tr>
<td>Habibian et al.</td>
<td>TPAMI 2017</td>
<td>20.0</td>
</tr>
</tbody>
</table>
State-of-the-art event retrieval

<table>
<thead>
<tr>
<th>Authors</th>
<th>Published</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Concept embedding</strong></td>
<td>ICMR 2014</td>
<td>6.4</td>
</tr>
<tr>
<td>Ye et al.</td>
<td>MM 2015</td>
<td>9.0</td>
</tr>
<tr>
<td>Chang et al.</td>
<td>IJCAI 2015</td>
<td>9.6</td>
</tr>
<tr>
<td>Mazloom et al.</td>
<td>ICMR 2015</td>
<td>11.9</td>
</tr>
<tr>
<td>Wu et al.</td>
<td>CVPR 2014</td>
<td>12.7</td>
</tr>
<tr>
<td>Jiang et al.</td>
<td>AAAI 2015</td>
<td>12.9</td>
</tr>
<tr>
<td><strong>Tag embedding</strong></td>
<td>TMM 2016</td>
<td>12.9</td>
</tr>
<tr>
<td>Liang et al.</td>
<td>MM 2015</td>
<td>18.3</td>
</tr>
<tr>
<td><strong>Video2vec embedding</strong></td>
<td>TPAMI 2017</td>
<td>20.0</td>
</tr>
</tbody>
</table>
State-of-the-art: event classification

<table>
<thead>
<tr>
<th>Authors</th>
<th>Published</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Habibian et al.</td>
<td>MM 2014</td>
<td>19.6</td>
</tr>
<tr>
<td>Nagel et al.</td>
<td>BMVC 2015</td>
<td>21.8</td>
</tr>
<tr>
<td>Li et al.</td>
<td>ICCV 2013</td>
<td>23.7</td>
</tr>
<tr>
<td>Tang et al.</td>
<td>CVPR 2012</td>
<td>26.8</td>
</tr>
<tr>
<td>Sun et al.</td>
<td>CVPR 2014</td>
<td>28.7</td>
</tr>
<tr>
<td>Chang et al.</td>
<td>MM 2015</td>
<td>30.9</td>
</tr>
<tr>
<td><strong>ImageNet-shuffle</strong></td>
<td>ICMR 2016</td>
<td>34.8</td>
</tr>
<tr>
<td><strong>Video2vec embedding</strong></td>
<td>TPAMI 2017</td>
<td>37.1</td>
</tr>
</tbody>
</table>
Conclusions

Event recognition without examples demands lingual representation

*Concept embedding* has too many limitations

*Tag embedding* is simple, yet surprisingly effective

*Video2vec*‘s descriptiveness & predictability is appealing