Real-World Applications of Activity Recognition

Sangmin Oh
Kitware

CVPR tutorial on 2014/06/23
Emerging Applications

Unconstrained Video Search

Aerial Video Analysis

Sports Video Analysis
Unconstrained Video Search
Lots of unconstrained video...
...find me activities I want

Task: Retrieve clips with activities of interest (e.g. “flash mob” or “birthday”)

Challenges:
Content variation across archive is huge
Content variation within activity is large
Metadata variations (frame size, clip length, bitrates, ...)
Archive size is large (150K+ clips)

Interaction
Unconstrained Video Search Datasets

**TRECVID Multimedia Event Detection (MED) Dataset**

- Evaluation data: Very large collection of web videos and detection of known event types.
- Available from a webpage (pending TRECVID participation): trecvid.nist.gov
- Complex events
  - 25 Test events (as of 2012, and increasing):
    - Wedding, changing a tire, woodworking project, parkour, townhall meeting, marriage proposal etc.
  - Full clips: Includes stitching, severe camera motion, temporal and spatial clutter, e.g., 1 Hour long.

**Columbia Consumer Video (CCV) dataset**

- Total 9317 videos (210 hours in total)
- Average length: 80 secs
- Complex events
  - 20 events
  - Wedding ceremony, wedding reception, biking, graduation, baseball, birthday, bird, playground etc
How does Random Look?

Random images from typical unconstrained videos
Flash Mob

Cykel
Flash mob
#1
Multiple Search Modes

Find videos similar to these examples
Query by video examples i.e., a set of videos

Find videos containing the following objects and scenes
Query by text e.g., People+ Dance + Dim light

Refine results with feedback

Large-Scale Multimedia Search Archive

Extracted Visual and Audio Features: Semantic + Low-level

DB  DB  DB  DB  DB  DB  DB
Examples of Video Features: Visual & Audio

Objects

- Sky
- People

Scene Attributes

- Indoor/outdoor
- Lighting
- Emotion
- Functions
- Materials
- Viewpoint

Low-level Visual Features

- Histogram of Oriented Gradients, Texture

Action: Crouching

Object: Car

Audio: Engine, Wind, Talk

Object: Tire

Scene: Urban, Street, Sunny, Outdoor, Building

Actions

- Indoor/outdoor
- Emotion
- Lighting
- Functions
- Materials
- Viewpoint

Videography

- Pan / Tile / Zoom
- Size of people
- Correlation of camera and FG motion

Low-level Audio Signatures (MFCCs)

- Engine
- Human Chat
- Outdoor
- Explosion
- Animal
- Water

Audio Events
Low-level Feature & Encoding

Local feature extraction

Quantization using Clustering Codebook

BoW Histogram

|sum of diffs to c(1)|…. |sum of diffs to c(n)|

Concatenate

Normalize

Difference Coding Vector

Dimension = K*D

Aggregating local descriptors into a compact image representation Jegou, Douze, Schimid, Perez, CVPR 2010.

Fisher Vectors for Fine-Grained Visual Categorization Perronnin, Sanchez, Akata, CVPR 2011

Large-scale Web Video Event Classification by use of Fisher Vectors, Sun, Navatia in WACV 2013
Difference coding method can achieve higher accuracy with lower computational demand. Most expensive step is quantization, and difference coding may require less number of quantizations are required due to reduced cluster centers.

Cost is potentially larger memory footprint.
Activities and Objects

Average Object detector responses on Wedding Videos (TRECVID MED dataset)

Image Courtesy of Greg Mori’s group at Simon Fraser Univ.

Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification

Li, Su, Xing, Fei-Fei, NIPS 2010
Videography Style Analysis

Combine a set of camera motion and related features into a “videography style descriptor”

Idea is for the style descriptor to capture some semantically meaningful things about how the video was taken

A Videography Analysis Framework for Video Retrieval and Summarization Oh, Li, Perera, Fu, BMVC 2012
Videography Styles

Example on Parkour video

Captures
• Background Motion
• Foreground Motion
• Correlation BG/FG
• Scale

• **Red**: background
• **Green**: foreground
• **White arrows**: Camera
Classifier Baseline Architecture

1. Feature Extraction (Visual, Audio)
   - Input Videos
     - Low-level Clip-level Pooling
     - Mid-level Clip-level Pooling
     - High-level Clip-level Pooling

2. Base Classifiers
   - SVM Linear, Nonlinear
   - SVM Linear, Nonlinear
   - SVM Linear, Nonlinear

3. Score Fusion
   - Mix-and-Match
     - Untrained Fusion (Average, GeoMean)
     - Learned Fusion

4. Complex Event Classification
   - Final Score

Baseline
- Clip-level Pooling + SVM

Video clip: \( x \)
Example: Board Trick
Segment-level features
Pooling
Avg, Max, Etc.

Support Vector Machine (SVMs)

Multimodal feature fusion for robust event detection in web videos Natarjan et al. CVPR 2012.
# Single Feature and Fusion Results

Different Events, (see TRECVID MED dataset for details)

<table>
<thead>
<tr>
<th>Base Classifier</th>
<th>E006</th>
<th>E007</th>
<th>E008</th>
<th>E009</th>
<th>E010</th>
<th>E011</th>
<th>E012</th>
<th>E013</th>
<th>E014</th>
<th>E015</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIST HIK</td>
<td>0.667</td>
<td>0.685</td>
<td>0.326</td>
<td>0.558</td>
<td>0.736</td>
<td>0.593</td>
<td>0.524</td>
<td>0.544</td>
<td>0.551</td>
<td>0.691</td>
</tr>
<tr>
<td>HoG3D Linear</td>
<td>0.645</td>
<td>0.775</td>
<td>0.341</td>
<td>0.484</td>
<td>0.724</td>
<td>0.700</td>
<td>0.452</td>
<td>0.519</td>
<td>0.590</td>
<td>0.728</td>
</tr>
<tr>
<td>HoG3D HIK</td>
<td>0.457</td>
<td>0.550</td>
<td>0.273</td>
<td>0.463</td>
<td>0.586</td>
<td>0.500</td>
<td>0.448</td>
<td>0.317</td>
<td>0.359</td>
<td>0.531</td>
</tr>
<tr>
<td>HoG3D NGD</td>
<td>0.462</td>
<td>0.468</td>
<td>0.258</td>
<td>0.474</td>
<td>0.529</td>
<td>0.479</td>
<td>0.352</td>
<td>0.327</td>
<td>0.321</td>
<td>0.531</td>
</tr>
<tr>
<td>ISA HIK</td>
<td>0.516</td>
<td>0.514</td>
<td>0.288</td>
<td>0.453</td>
<td>0.540</td>
<td>0.536</td>
<td>0.424</td>
<td>0.317</td>
<td>0.333</td>
<td>0.506</td>
</tr>
<tr>
<td>CSIFT Linear</td>
<td>0.704</td>
<td>0.649</td>
<td>0.364</td>
<td>0.453</td>
<td>0.747</td>
<td>0.679</td>
<td>0.463</td>
<td>0.683</td>
<td>0.526</td>
<td>0.654</td>
</tr>
<tr>
<td>CSIFT HIK</td>
<td>0.468</td>
<td>0.414</td>
<td>0.197</td>
<td>0.368</td>
<td>0.540</td>
<td>0.493</td>
<td>0.359</td>
<td>0.413</td>
<td>0.372</td>
<td>0.506</td>
</tr>
<tr>
<td>CSIFT NGD</td>
<td>0.495</td>
<td>0.405</td>
<td>0.205</td>
<td>0.379</td>
<td>0.540</td>
<td>0.479</td>
<td>0.368</td>
<td>0.423</td>
<td>0.359</td>
<td>0.519</td>
</tr>
<tr>
<td>TCH Linear</td>
<td>0.774</td>
<td>0.676</td>
<td>0.447</td>
<td>0.474</td>
<td>0.805</td>
<td>0.793</td>
<td>0.489</td>
<td>0.712</td>
<td>0.564</td>
<td>0.753</td>
</tr>
<tr>
<td>TCH HIK</td>
<td>0.532</td>
<td>0.532</td>
<td>0.273</td>
<td>0.389</td>
<td>0.575</td>
<td>0.543</td>
<td>0.398</td>
<td>0.404</td>
<td>0.397</td>
<td>0.593</td>
</tr>
<tr>
<td>TCH NGD</td>
<td>0.532</td>
<td>0.477</td>
<td>0.250</td>
<td>0.400</td>
<td>0.621</td>
<td>0.536</td>
<td>0.420</td>
<td>0.404</td>
<td>0.423</td>
<td>0.580</td>
</tr>
<tr>
<td>OB AVG, Linear (L0)</td>
<td>0.645</td>
<td>0.550</td>
<td>0.311</td>
<td>0.432</td>
<td>0.644</td>
<td>0.600</td>
<td>0.519</td>
<td>0.625</td>
<td>0.590</td>
<td>0.617</td>
</tr>
<tr>
<td>OB MAX, Linear (L0)</td>
<td>0.597</td>
<td>0.541</td>
<td>0.379</td>
<td>0.442</td>
<td>0.632</td>
<td>0.500</td>
<td>0.554</td>
<td>0.558</td>
<td>0.474</td>
<td>0.630</td>
</tr>
<tr>
<td>OB, LSVM (L1)</td>
<td>0.570</td>
<td>0.577</td>
<td>0.280</td>
<td>0.568</td>
<td>0.621</td>
<td>0.586</td>
<td>0.442</td>
<td>0.538</td>
<td>0.500</td>
<td>0.605</td>
</tr>
<tr>
<td>OB Avg, HIK</td>
<td>0.532</td>
<td>0.505</td>
<td>0.250</td>
<td>0.411</td>
<td>0.575</td>
<td>0.550</td>
<td>0.442</td>
<td>0.375</td>
<td>0.436</td>
<td>0.568</td>
</tr>
<tr>
<td>OB Max, HIK</td>
<td>0.516</td>
<td>0.477</td>
<td>0.250</td>
<td>0.387</td>
<td>0.529</td>
<td>0.457</td>
<td>0.420</td>
<td>0.442</td>
<td>0.385</td>
<td>0.519</td>
</tr>
<tr>
<td>SUN09 MKL</td>
<td>0.441</td>
<td>0.351</td>
<td>0.205</td>
<td>0.337</td>
<td>0.483</td>
<td>0.507</td>
<td>0.355</td>
<td>0.337</td>
<td>0.321</td>
<td>0.506</td>
</tr>
<tr>
<td>MFCCs Linear</td>
<td>0.548</td>
<td>0.782</td>
<td>0.545</td>
<td>0.681</td>
<td>0.814</td>
<td>0.761</td>
<td>0.645</td>
<td>0.709</td>
<td>0.346</td>
<td>0.667</td>
</tr>
<tr>
<td>MFCCs HIK</td>
<td>0.446</td>
<td>0.618</td>
<td>0.424</td>
<td>0.564</td>
<td>0.686</td>
<td>0.739</td>
<td>0.584</td>
<td>0.593</td>
<td>0.295</td>
<td>0.654</td>
</tr>
<tr>
<td>MFCCs NGD</td>
<td>0.462</td>
<td>0.673</td>
<td>0.409</td>
<td>0.500</td>
<td>0.698</td>
<td>0.696</td>
<td>0.567</td>
<td>0.631</td>
<td>0.372</td>
<td>0.628</td>
</tr>
<tr>
<td>ASM186 HIK</td>
<td>0.422</td>
<td>0.561</td>
<td>0.470</td>
<td>0.553</td>
<td>0.744</td>
<td>0.618</td>
<td>0.607</td>
<td>0.602</td>
<td>0.289</td>
<td>0.553</td>
</tr>
<tr>
<td>ASM64 HIK</td>
<td>0.438</td>
<td>0.607</td>
<td>0.523</td>
<td>0.628</td>
<td>0.733</td>
<td>0.669</td>
<td>0.620</td>
<td>0.670</td>
<td>0.303</td>
<td>0.618</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fusion Model</th>
<th>E006</th>
<th>E007</th>
<th>E008</th>
<th>E009</th>
<th>E010</th>
<th>E011</th>
<th>E012</th>
<th>E013</th>
<th>E014</th>
<th>E015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0.265</td>
<td>0.318</td>
<td>0.212</td>
<td>0.234</td>
<td>0.339</td>
<td>0.426</td>
<td>0.298</td>
<td>0.262</td>
<td>0.184</td>
<td>0.447</td>
</tr>
<tr>
<td>GeoMean</td>
<td>0.292</td>
<td>0.290</td>
<td>0.189</td>
<td>0.266</td>
<td>0.430</td>
<td>0.404</td>
<td>0.281</td>
<td>0.252</td>
<td>0.224</td>
<td>0.461</td>
</tr>
<tr>
<td>MFoM</td>
<td>0.324</td>
<td>0.299</td>
<td>0.197</td>
<td>0.287</td>
<td>0.430</td>
<td>0.419</td>
<td>0.329</td>
<td>0.262</td>
<td>0.237</td>
<td>0.487</td>
</tr>
<tr>
<td>LEF</td>
<td>0.265</td>
<td>0.318</td>
<td>0.197</td>
<td>0.245</td>
<td>0.384</td>
<td>0.412</td>
<td>0.285</td>
<td>0.233</td>
<td>0.211</td>
<td>0.461</td>
</tr>
<tr>
<td>Best Base - Best Fusion</td>
<td>0.157</td>
<td>0.061</td>
<td>0.008</td>
<td>0.103</td>
<td>0.099</td>
<td>0.053</td>
<td>0.071</td>
<td>0.084</td>
<td>0.105</td>
<td>0.059</td>
</tr>
</tbody>
</table>

* Lower number indicates higher accuracy

Results from **Multimedia event detection with multimodal feature fusion and temporal concept localization** Oh, McCloskey, Kim, at al. Machine Vision and Applications 25(1), 2014.
Event Structure Learning

Events have certain structures consisting of salient parts and non-important regions. How do we exploit and learn these?

Example: “Making a sandwich”
Spatio-Temporal Weakly Supervised Learning

- **Weakly supervised learning formulation**
  - How do we identify important and salient segments from videos belonging to same events?
  - Can this be done implicitly or explicitly?
  - What should be the granularity in time and feature space which will work?
Mid-level: Frame Clusters
Learning Structure Implicitly using Topic-based Pooling


Scene aligned pooling for complex video recognition, Cao, Mu, Natsev, Chang, Hua, Smith, in ECCV 2012.
Recognition by Composition: Latent Temporal Part-based Learning

Explicitly Searches for
- Representative Segments
- Best Feature Combinations
- Best ‘Hidden’ segment Types

<table>
<thead>
<tr>
<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>E6</td>
<td>0.54</td>
<td>1.97</td>
<td>2.25</td>
<td>4.38</td>
<td>2.77</td>
<td>2.34</td>
<td>6.08</td>
<td>4.73</td>
<td>5.73</td>
</tr>
<tr>
<td>E7</td>
<td>0.35</td>
<td>1.25</td>
<td>0.76</td>
<td>0.92</td>
<td>2.11</td>
<td>1.33</td>
<td>2.87</td>
<td>2.26</td>
<td>4.81</td>
</tr>
<tr>
<td>E8</td>
<td>0.42</td>
<td>6.48</td>
<td>8.30</td>
<td>15.29</td>
<td>25.48</td>
<td>10.30</td>
<td>20.75</td>
<td>22.99</td>
<td>35.82</td>
</tr>
<tr>
<td>E9</td>
<td>0.26</td>
<td>2.15</td>
<td>1.95</td>
<td>2.04</td>
<td>4.14</td>
<td>1.79</td>
<td>6.25</td>
<td>7.61</td>
<td>8.38</td>
</tr>
<tr>
<td>E10</td>
<td>0.25</td>
<td>0.81</td>
<td>0.74</td>
<td>0.74</td>
<td>1.03</td>
<td>0.76</td>
<td>1.43</td>
<td>1.34</td>
<td>2.12</td>
</tr>
<tr>
<td>E11</td>
<td>0.43</td>
<td>1.10</td>
<td>1.48</td>
<td>0.84</td>
<td>1.93</td>
<td>1.41</td>
<td>2.29</td>
<td>2.65</td>
<td>4.65</td>
</tr>
<tr>
<td>E12</td>
<td>0.58</td>
<td>5.83</td>
<td>2.65</td>
<td>4.03</td>
<td>7.06</td>
<td>5.71</td>
<td>8.44</td>
<td>8.70</td>
<td>10.99</td>
</tr>
<tr>
<td>E13</td>
<td>0.32</td>
<td>2.58</td>
<td>2.05</td>
<td>3.04</td>
<td>10.38</td>
<td>2.57</td>
<td>9.44</td>
<td>10.43</td>
<td>13.11</td>
</tr>
<tr>
<td>E14</td>
<td>0.27</td>
<td>1.18</td>
<td>4.39</td>
<td>10.88</td>
<td>6.69</td>
<td>4.58</td>
<td>10.00</td>
<td>11.89</td>
<td>23.32</td>
</tr>
<tr>
<td>E15</td>
<td>0.26</td>
<td>0.92</td>
<td>0.61</td>
<td>5.48</td>
<td>1.21</td>
<td>1.09</td>
<td>2.49</td>
<td>2.4</td>
<td>3.29</td>
</tr>
<tr>
<td>mAP</td>
<td>0.37</td>
<td>2.43</td>
<td>2.52</td>
<td>4.77</td>
<td>6.28</td>
<td>3.19</td>
<td>7.00</td>
<td>7.50</td>
<td>11.22</td>
</tr>
</tbody>
</table>
It is possible to get a high quality set of matching videos from large archive

Precision @ 32 = 97% as shown; AP = 74.3%; archive contains 26K videos including ~100 true positives
Top 30 for “Vehicle Tire Change”
100 positive training examples used

- It is possible to get a high quality set of matching videos from large archive

  Precision @ 32 = 84% as shown; AP = 52.6%; archive contains 26K videos including ~100 true positives
Activity Recognition in Aerial Videos
**Video from Sky**

**Characteristics**
- Large Images/Videos
- Mostly vertical point of view
- Moving camera
- Small objects
- Lighting/Occlusion by nature
- Can have substantial scale changes

**Application domains**
- Disaster relief
- Emergency responder
- Broadcasting
- Traffic surveying/control
- Business Intelligence
- Security
- Military
Sensors: FMV and WAMI

**Full Motion Video (FMV)**
- Mostly single camera
- Moderate resolution
- User control
- Substantial camera motion

**Wide Area Motion Imagery**
- Multiple camera array
- Image stitching
- Very large image format
- Fairly good stabilization to point to certain area
Wright-Patterson Air Force Base (WPAFB) 2009 Dataset

- Six cameras with ortho-rectified (stitched and geo-registered) imagery
- Image size: > 20K x 20K pixels
- GSD: 25 cm/pixel
- Frame rate: ~1.25Hz
- NITF file format with encoded sensor metadata
- 21 minutes (1,537 frames) of video
- **14 minutes (1,025 frames) with over 18K ground truth tracks**
- Publicly released by Air Force Research Lab (AFRL) SDMS

WPAFB Dataset: Track Ground Truth

6,500+ ground-truth tracks in 7 minutes
Large-Scale Real-time Long-term Tracking

Real-time Multi-Target Tracking at 210 Megapixels/second in Wide Area Motion Imagery, Basharat, Turek, Xu, Atkins, Stoup, Fieldhouse, Tunison, Hoogs, in WACV 2014
Large-Scale Real-time Long-term Tracking

Latest Unlinked Tracklet
Linked Track

- Tracks from an Area of Interest (AOI) processed as a single tile
- 6 min long track
- Includes inter-tile linking

*Real-time Multi-Target Tracking at 210 Megapixels/second in Wide Area Motion Imagery*, Basharat, Turek, Xu, Atkins, Stoup, Fieldhouse, Tunison, Hoogs, in WACV 2014
# Events & Actions

<table>
<thead>
<tr>
<th>Articulated Motion (Sub-entity)</th>
<th>Single-entity</th>
<th>Two-entity</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exploding</td>
<td>Exploding</td>
<td>Exploding</td>
<td>Speaking to crowds</td>
</tr>
<tr>
<td>Burning</td>
<td>Burning</td>
<td>Burning</td>
<td></td>
</tr>
<tr>
<td>Digging</td>
<td>Digging</td>
<td>Kissing</td>
<td></td>
</tr>
<tr>
<td>Picking up</td>
<td>Picking up</td>
<td>Exchanging</td>
<td></td>
</tr>
<tr>
<td>Throwing</td>
<td>Throwing</td>
<td>objects</td>
<td></td>
</tr>
<tr>
<td>Carrying</td>
<td>Carrying</td>
<td>Kicking</td>
<td></td>
</tr>
<tr>
<td>Shooting</td>
<td>Shooting</td>
<td>Carrying</td>
<td></td>
</tr>
<tr>
<td>Launching</td>
<td>Launching</td>
<td>together</td>
<td></td>
</tr>
<tr>
<td>Limping</td>
<td>Limping</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kicking</td>
<td>Kicking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td>Smoking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gesturing</td>
<td>Gesturing</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relative Motion (Track-level)</th>
<th>Single-entity</th>
<th>Two-entity</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>Walking</td>
<td>Getting in/out</td>
<td>Convoy</td>
</tr>
<tr>
<td>Running</td>
<td>Running</td>
<td>Passing off</td>
<td>Receiving line</td>
</tr>
<tr>
<td>Loitering</td>
<td>Loitering</td>
<td>Picking up</td>
<td>Troop formation</td>
</tr>
<tr>
<td>Starting</td>
<td>Starting</td>
<td>Overtaking or passing</td>
<td></td>
</tr>
<tr>
<td>Turning</td>
<td>Turning</td>
<td>Overtaking</td>
<td></td>
</tr>
<tr>
<td>U-turn</td>
<td>U-turn</td>
<td>Overtaking</td>
<td></td>
</tr>
<tr>
<td>Stopping</td>
<td>Stopping</td>
<td>Overtaking</td>
<td></td>
</tr>
<tr>
<td>Aimless Driving</td>
<td>Aimless Driving</td>
<td>Overtaking</td>
<td></td>
</tr>
<tr>
<td>Accelerating</td>
<td>Accelerating</td>
<td>Overtaking</td>
<td></td>
</tr>
<tr>
<td>Decelerating</td>
<td>Decelerating</td>
<td>Overtaking</td>
<td></td>
</tr>
<tr>
<td>Following</td>
<td>Following</td>
<td>Maintaining</td>
<td></td>
</tr>
<tr>
<td>Meeting</td>
<td>Meeting</td>
<td>distance</td>
<td></td>
</tr>
<tr>
<td>Gathering</td>
<td>Gathering</td>
<td>Forming</td>
<td></td>
</tr>
<tr>
<td>Moving as a group</td>
<td>Moving as a group</td>
<td>Forming</td>
<td></td>
</tr>
<tr>
<td>Dispersing</td>
<td>Dispersing</td>
<td>convoys</td>
<td></td>
</tr>
</tbody>
</table>

### Data Requirements:
- **Low Resolution:** possible by analyzing track-level information
- **High Resolution:** requires detailed pixel information
Continuous Visual Event Recognition (CVER)

Common Architecture

- Foreground motion detection, e.g., tracking etc.
- Temporal segmentation, e.g., regular/variable units
- Classification, e.g., 1-vs-All, multi-way etc.
- Upper bound determined by weakest among above

Lots of blank intervals/space are challenging to optimize precision and recall
At low resolution, many actions look very similar

<table>
<thead>
<tr>
<th>Action</th>
<th>Image 1</th>
<th>Image 2</th>
<th>Image 3</th>
<th>Image 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>standing</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>digging</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>walking</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>carrying</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>running</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
</tbody>
</table>
Event Models & Features

**Pixel-based Features**
- spatio-temporal histogram of gradients [6][12]
- optical flow [2][6][9]
- Feature point descriptor: SIFT [10][11][15]
- histogram of optical flow [6]

**Macro Features**
- sensor metadata: gsd, pointing angles, etc.
- bounding box: area, aspect ratio, etc.
- track level: speed, delta heading, curvature, etc. [1][12][13][8]
- periodicity of self-similarity matrices [14]

**Models & Classification**
- BoW + SVM [4][11][16]
- Dynamic Bayesian networks [12][13]
- objects interaction modeling [1][7][8]

---

Feature point descriptor:
- SIFT [10][11][15]
References


References (cont’d)

11. SF Chang, J He, YG Jiang, CW Ngo, A Yanagawa, Columbia University/VIRED-CityU/IRIT TRECVID2008 high-level feature extraction and interactive video search


13. Zhi Zeng and Qiang Ji, Knowledge Based Activity Recognition with Dynamic Bayesian Network, ECCV 2010


Event Detections on WPAFB
Event Detections on FMV dataset
Activity-based Scene Understanding

**Objective:** Recognize stationary scene elements based on surrounding pedestrian and vehicle behaviors, as opposed to appearance features

Where are more scene elements similar to these?
Key Challenges (Multi-Modal Behaviors)

Different scene element instances can have significantly different behaviors

Roadways
- Many Modes
- Few Modes

Intersection
- Many Modes
- Few Modes

Event Legend
- Vehicle Driving
- Vehicle Turning
- Vehicle Starting
- Vehicle Stopping
- Person Walking

Pyramid Coding for Functional Scene Element Recognition in Video Scenes, Swears, Boyer, and Hoogs, in ICCV 2013
Features: Object, Motion, Statistics

**Objective:** Extract activity behavior descriptors using automatically computed tracks

*Pyramid Coding for Functional Scene Element Recognition in Video Scenes*, Swears, Boyer, and Hoogs, in ICCV 2013
Activity-based Scene Understanding

True Evaluated Scene Elements

Pyramid Coding, MRF Labels

Buildings
Intersection
Cross-walk
Roadway
Sidewalk
Doorway

Overall

Pyramid Coding for Functional Scene Element Recognition in Video Scenes, Swears, Boyer, and Hoogs, in ICCV 2013
Where are we and heading to?

We are somewhere here

What can we do?

FMV
- Event detection
- Event-based video indexing

WAMI
- Event detection
- Normalcy models & anomaly detection
- Complex activity recognition
Sports Video Analytics
Sports Analytics

Application domains

• Player Tracking
• Event / Strategy recognition
• Event-based indexing
• (Semi) Automatic camera capture control
• Summarization
Early Work: Intille and Bobick

- Assumed Perfect Tracking information and role recognition
- Bayesian networks to agglomerate evidence and infer team playing strategies
- Limitation: Sports videos are chaotic, and we never get perfect features!

**Recognizing Planned Multiperson Action**, Intille and Bobick, Computer Vision and Image Understanding 81, 2001
Today: Towards Automated Sports Broadcasting

How do we help a single or a small number of crew(s) to manage the entire broadcasting feed?

Group Motion Prediction for Camera Control

Example 1: Interception & Goal keeping - Result

Backdoor play (through pass) (soccer)

Compute motion fields based on player motions. Then, find convergence points to predict where ball (and players) are moving to.

Prediction-based Video Re-targeting

Simulation of camera control: Example 1

Original Video Feed among many camera feeds

Automatically computed Re-targetted video

Predicting Wins based on detailed trajectory analysis

Sweet-Spot: Using Spatiotemporal Data to Discover and Predict Shots in Tennis, Wei, Lucey, Morgan, Sridharan, in MIT Sloan Sports Analytics Conference
Predicting Wins based on detailed trajectory analysis

Sweet-Spot: Using Spatiotemporal Data to Discover and Predict Shots in Tennis, Wei, Lucey, Morgan, Sridharan, in MIT Sloan Sports Analytics Conference
Swimming Video Analysis

Analyzing a Large Collection of Archived Swimming Videos

Calibration

Swimmer Location

Part Tracking


Understanding and Analyzing a large Collection of Archived Swimming Videos,
Sha, Lucey, Morgan, Sridharan, Pease, in in WACV 2014
Complex Play Recognition with Imperfect Tracks

Which play is being run? How soon can we tell?

Spatial variability

Fragmented and partial tracks

Partial Temporal Ordering

Complex object

Camera Motion

Active Deception
Play Taxonomy

American Football Plays

Pass
- Option
- Rollout
- Short
- Deep

Run
- Left
- Middle
- Right

Learning and Recognizing Complex Multi-Agent Activities with Applications to American Football Plays, Swears and Hoogs in WACV 2012
Robust Play Recognition against Track Fragmentation

Track Normalization

Learn Non-Stationary HMM using positions and speeds from tracks

Tracker ID is not important anymore!

Accuracy Improves with More observations

Learning and Recognizing Complex Multi-Agent Activities with Applications to American Football Plays, Swars and Hoogs in WACV 2012
Summary

Unconstrained Video Search

Aerial Video Analysis

Sports Video Analysis