Looking at Egocentric Video

Cameras

Users

Videos
Steve Mann is Online Since 1980

- Steve’s live camera online – who watched?
- Looking through other people’s eyes – will people share what they see?
- Will we have “channels” selecting among cameras?
Temporal Segmentation of Egocentric Videos

Yair Poleg  Chetan Arora  Shmuel Peleg

To appear in CVPR`14, June 2014
Life-logging Videos

• Video is always on.
  – No “record” button.
  – No viewfinder.

• Infinite, continuous, unstructured video
  – Jumpy - head always moving
  – Boring
  – Hard to search and browse
Research on Egocentric Video

• Activity Recognition [Alireza CVPR`11]
  – Minutes (preparing tuna sandwich)
  – Hands, **object recognition**, gaze
  – Limited to controlled environments
  – Limited # of activities

• Action recognition [Kitani CVPR`11]
  – Seconds (jumping up)
  – Wearer’s head motion (Flow + Frequency)
  – Unsupervised clustering, Short term actions
  – Temporal over-segmentation
Research on Egocentric Video

• Social Interactions [Alireza CVPR`12]
  – Who’s looking where?
  – Who’s talking to me?
  – Gaze/Look-at

• Summarization [Grauman CVPR`13]
  – Make long story short.
  – Recognizing Important people.
  – Recognizing objects/places.
Divide Long Video to Chapters

• Possible data that can be used:
  – Interactions with recognized objects & faces
  – Context: place & scene recognition
  – Wearer’s Motion

• Example: Use motion to divide video to 3 classes:
  – Stationary, Walking, Riding
  – Wearer’s speed is enough

• Why not GPS or Ego-motion Estimation?
  – GPS works outdoors, but fails indoors
  – Ego-motion fails in long videos (Point tracking fails)
Stationary vs. In Transit

Estimate Optical Flow in cells of a fixed grid

Normal “transit” is looking forward

Grid of 10x5 Blocks

Expected Optical Flow
Transit Optical Flow

Optical Flow in time $t$

Expected Optical Flow

Measured Optical Flow
Transit Optical Flow

Optical Flow in time $t+1$

Expected Optical Flow

Measured Optical Flow
Why This Optical Flow?

Examine over time the optical flow in one cell. Head rotations dominate optical flow.
Integration Over Time of Optical Flow

- Use long-term integration for classification
  - Different colors = different cells
  - Integration removes head motion
Integration Over 3500 first frames
Proposed Motion Classes

Input Video

Stationary
- Static
- Moving Head
  - Sitting
  - Standing

Transit
- Open View
- Box
  - Walking
  - Wheels
  - Car
  - Bus
Motion Patterns ➔ Features

- Motion Vectors
  - Smooth Cumulative Displacement Curves
Motion Patterns ➔ Features

• Motion Vectors
  – Smooth Cumulative Displacement Curves
  – Find slope of curves
  – Note: Same as blurring the original Optical Flow

• Other approaches:
  – Piecewise linear approx
  – Down/Up sampling with various interpolations
  – Whatever segmentation approach..
Motion Based Features

- **Radial Projection Response**
  - Project flow onto template
  - Count OK cells

- **Flow Magnitude Clusters**

- **Statistical Information:**
  - Number of blocks with valid Optical Flow
  - Avg & Stddev of flow magnitudes
Training Binary Classifiers

• Dataset:
  – 140 sequences, 65hrs, ~3.5M frames
  – All labeled into 7 classes by students

• Training Set:
  – Randomly pick a sequences until we cover 12K samples per class
  – Training sequences are excluded from the test

• Test Set: Unseen sequences
Confusion Matrix – Leaf Nodes

### Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>Walking</th>
<th>Car</th>
<th>Standing</th>
<th>Bus</th>
<th>Wheels</th>
<th>Sitting</th>
<th>Static</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>83%</td>
<td>0%</td>
<td>6%</td>
<td>6%</td>
<td>4%</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>Car</td>
<td>1%</td>
<td>74%</td>
<td>3%</td>
<td>15%</td>
<td>0%</td>
<td>3%</td>
<td>4%</td>
</tr>
<tr>
<td>Standing</td>
<td>14%</td>
<td>2%</td>
<td>47%</td>
<td>4%</td>
<td>0%</td>
<td>31%</td>
<td>2%</td>
</tr>
<tr>
<td>Bus</td>
<td>3%</td>
<td>19%</td>
<td>27%</td>
<td>43%</td>
<td>0%</td>
<td>7%</td>
<td>1%</td>
</tr>
<tr>
<td>Wheels</td>
<td>9%</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
<td>86%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Sitting</td>
<td>3%</td>
<td>1%</td>
<td>22%</td>
<td>1%</td>
<td>0%</td>
<td>62%</td>
<td>10%</td>
</tr>
<tr>
<td>Static</td>
<td>0%</td>
<td>1%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
<td>97%</td>
</tr>
</tbody>
</table>
### Confusion Matrix – Leaf Nodes

#### Input Video

- **Stationary**
  - **Static**
  - **Moving Head**
- **Transit**
  - **Open View**
  - **Box**
  - **Sitting**
  - **Standing**
  - **Walking**
  - **Wheels**
  - **Car**
  - **Bus**

#### Table

<table>
<thead>
<tr>
<th></th>
<th>Walking</th>
<th>Car</th>
<th>Standing</th>
<th>Bus</th>
<th>Wheels</th>
<th>Sitting</th>
<th>Static</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>83%</td>
<td>0%</td>
<td>6%</td>
<td>6%</td>
<td>4%</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>Car</td>
<td>1%</td>
<td><strong>74%</strong></td>
<td>3%</td>
<td>15%</td>
<td>0%</td>
<td>3%</td>
<td>4%</td>
</tr>
<tr>
<td>Standing</td>
<td>14%</td>
<td>2%</td>
<td><strong>47%</strong></td>
<td>4%</td>
<td>0%</td>
<td><strong>31%</strong></td>
<td>2%</td>
</tr>
<tr>
<td>Bus</td>
<td>3%</td>
<td>19%</td>
<td>27%</td>
<td><strong>43%</strong></td>
<td>0%</td>
<td>7%</td>
<td>1%</td>
</tr>
<tr>
<td>Wheels</td>
<td>9%</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
<td><strong>86%</strong></td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Sitting</td>
<td>3%</td>
<td>1%</td>
<td>22%</td>
<td>1%</td>
<td>0%</td>
<td><strong>62%</strong></td>
<td>10%</td>
</tr>
<tr>
<td>Static</td>
<td>0%</td>
<td>1%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
<td><strong>97%</strong></td>
</tr>
</tbody>
</table>
Confusion Matrix – Leaf Nodes

![Confusion Matrix Diagram]

<table>
<thead>
<tr>
<th></th>
<th>Walking</th>
<th>Car</th>
<th>Standing</th>
<th>Bus</th>
<th>Wheels</th>
<th>Sitting</th>
<th>Static</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>83%</td>
<td>0%</td>
<td>6%</td>
<td>6%</td>
<td>4%</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>Car</td>
<td>1%</td>
<td>74%</td>
<td>3%</td>
<td>15%</td>
<td>0%</td>
<td>3%</td>
<td>4%</td>
</tr>
<tr>
<td>Standing</td>
<td>14%</td>
<td>2%</td>
<td>47%</td>
<td>4%</td>
<td>0%</td>
<td>31%</td>
<td>2%</td>
</tr>
<tr>
<td>Bus</td>
<td>3%</td>
<td>19%</td>
<td>27%</td>
<td>43%</td>
<td>0%</td>
<td>7%</td>
<td>1%</td>
</tr>
<tr>
<td>Wheels</td>
<td>9%</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
<td>86%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Sitting</td>
<td>3%</td>
<td>1%</td>
<td>22%</td>
<td>1%</td>
<td>0%</td>
<td>62%</td>
<td>10%</td>
</tr>
<tr>
<td>Static</td>
<td>0%</td>
<td>1%</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
<td>97%</td>
</tr>
</tbody>
</table>
Confusion Matrix – Inner Nodes

```
Input Video
    /\ Stationary       Transit
       /     \             /     \
  Static    Moving Head  Open View
       /     \             /     \
 Sitting   Standing     Walking
       /     \             /     \
 Wheels    Car           Bus
```

<table>
<thead>
<tr>
<th>Class Label</th>
<th>Accuracy</th>
<th># Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static-Moving</td>
<td>91%</td>
<td>1083115</td>
</tr>
<tr>
<td>Sitting-Standing</td>
<td>82%</td>
<td>1036217</td>
</tr>
<tr>
<td>Box-Open</td>
<td>87%</td>
<td>1197623</td>
</tr>
<tr>
<td>Car-Bus</td>
<td>76%</td>
<td>228108</td>
</tr>
<tr>
<td>Walking-Wheels</td>
<td>82%</td>
<td>969515</td>
</tr>
</tbody>
</table>
Failure Cases

• Mixed Activity – Standing in line
  – Naïve smoothing has its limitations..

• Relative Velocity
  – Train enters train station

• Disney’s Open Train
Wisdom of the Crowd in Egocentric Video Curation

Yedid Hoshen     Gil Ben-Artzi     Shmuel Peleg

The 3rd Workshop on Egocentric Vision 2014 at CVPR 2014, June 2014
Summarize Popular Events

- Several videos taken in popular events
  - Concerts, Lectures
- Curate a single video taking the best from each
- Avoid:
  - Personal moments (messaging, looking sideways)
  - Sharp head motions. Blurry images.
- Include:
  - Object of Interest
  - Sharp, stable frames
Video Quality Criteria

- Quality for individual frames:
  - Stability
  - Sharpness
  - Object in center
- Scene Popularity (wisdom of the crowd)
  - Exclude messaging etc.
- Smooth Transitions
- Cinematography
and more...
Existing Solutions

- **Video Mashup, Saini et al. (2012)**
  
  Assumes all frames are interesting

- **Social Cameras, Arev et al. (SIGGRAPH 2014)**
  
  Requires 3D reconstruction of the scene and 3D camera pose estimation.
Our Approach

• Frame Quality
  • Stability
  • Sharpness

• Scene Popularity
  • How many people are looking at same scene?

• Smooth Transitions
  • Large overlap in transition between cameras
Finding Regions of Interest

- Calculate $C_b, C_r$ color histogram of each frame
- Cluster histograms into ROI of each video
  - 1 cluster in concert, course
  - 2 clusters in a 3-person meeting
- Improve ROI assignment to frames using HMM
- Use chromo-temporal criterion to match ROI in different videos, using Normalized Cut
  - Due to view changes, color works better than shape descriptors
Popularity Measure

- How many users look at ROI at this time
Results
Results

https://www.youtube.com/channel/UCce6USxoqtBh9-MRRnnlQwg/videos